

# Journal of Experimental Psychology: General

## Arbitrary Fairness in Reward and Punishments

Ellen R. K. Evers, Michael O'Donnell, and Yoel Inbar

Online First Publication, November 28, 2022. <https://dx.doi.org/10.1037/xge0001300>

### CITATION

Evers, E. R. K., O'Donnell, M., & Inbar, Y. (2022, November 28). Arbitrary Fairness in Reward and Punishments. *Journal of Experimental Psychology: General*. Advance online publication. <https://dx.doi.org/10.1037/xge0001300>

# Arbitrary Fairness in Reward and Punishments

Ellen R. K. Evers<sup>1</sup>, Michael O'Donnell<sup>2</sup>, and Yoel Inbar<sup>3</sup>

<sup>1</sup> Department of Marketing, Haas School of Business, University of California, Berkeley

<sup>2</sup> Department of Marketing, McDonough School of Business, Georgetown University

<sup>3</sup> Department of Psychology, University of Toronto Scarborough

People have a strong preference for fairness. For many, fairness means equal reward and punishments for equal efforts and offences. However, this belief does not specify the units in which equality should be expressed. We show that people generally fail to take the interchangeability of units into account when judging and assigning fair punishments and reward. Therefore, judgments about and distributions of resources are strongly influenced by arbitrary decisions about which unit to express them in. For example, if points represent different monetary values for different recipients, people attempt to distribute money equally if money is salient but attempt to distribute points equally if points are salient. Because beliefs about fairness are a fundamental principle in many domains, the implications of these findings are broad. Essentially any distribution of outcomes can be made to appear more or less fair by changing the units these outcomes are expressed in.

*Keywords:* equity, fairness, heuristics, punishments and rewards, simple beliefs

*Supplemental materials:* <https://doi.org/10.1037/xge0001300.supp>

People care about fairness. In their interpersonal relationships, people strive to fairly balance costs and benefits (Fiske, 1991). Even when interacting with strangers in economic games, people are reluctant to maximize their earnings if doing so reduces fairness (Bolton et al., 1998). People not only have strong beliefs about what constitutes fair outcomes for themselves, but about outcomes experienced by others. People imbue judgments about fairness with moral beliefs, making preferences over allocations of goods and resources different from typical everyday preferences, which lack this moral component. In this way, judgments about fairness may take on characteristics of “moral mandates” (Skitka, 2010), which are more strongly held and more motivating than other types of beliefs. For instance, when people merely observe

an interaction that they perceive to be unfair, they are often willing to give up substantial amounts of money just to punish the person they perceive to be acting unfairly (e.g., Marlowe et al., 2008). Similar behavior can be observed in primates (Brosnan & de Waal, 2003), suggesting that a preference for fairness may be innate.

A large literature investigates beliefs about the fair allocation of outcomes. In most cases, the degree to which outcomes are distributed equitably (e.g., input linearly mapping onto output) plays a large role in whether people believe those outcomes to have been distributed fairly or not (e.g., Adams, 1965; Walster et al., 1978). When effort is similar or undefined, people most often believe equal outcomes represent fair distributions and exhibit an aversion to inequality (e.g., Charness & Rabin, 2002; Fehr & Schmidt, 1999; Rabin, 1993). These beliefs are so strong that people regularly reject outcome-maximizing but unequal ultimatum game offers, even if these offers are worth several days' wages (Henrich et al., 2005).

Although it is descriptively true that people often prefer outcomes with equal allocations of resources, this statement elides an important consideration: equal outcomes in which unit of expression? Many outcomes can be expressed in monetary terms as well as in other numeric units. For example, suppose a firm wants to give bonuses to a group of workers who earn different hourly wages. The firm might give every worker the same monetary bonus, even though the same amount represents different numbers of hours worked for different workers. Or the firm might give every worker an equal amount of paid time off, even though the same amount of time off represents different wages for different workers. Moreover, employees may judge equivalent allocations as more or less fair depending on which unit (hours or dollars) they are expressed in. Here, we study how people form impressions of what they perceive as fair distributions under conditions of interchangeability, that is,

---

Ellen R. K. Evers  <https://orcid.org/0000-0002-8667-3083>

Michael O'Donnell  <https://orcid.org/0000-0002-8953-3609>

Yoel Inbar  <https://orcid.org/0000-0002-3176-3727>

We thank Ariana Munroe, Cjache Kang and members of the preferences lab for their help running Study 4. This research was supported in part by the Berkeley X-Lab and the Fetzer Franklin fund of the John E. Fetzer Memorial Trust. Parts of this research have been presented at the Wharton Decision Processes Colloquium, the Society of Judgment and Decision Making Conference, and the Subjective Probability, Utility, and Decision Making Conference. We thank audience members at these seminars for their feedback and questions. All materials, data, and preregistrations can be found at <https://osf.io/9tcbx/> (Evers et al., 2022).

Correspondence concerning this article should be addressed to Ellen R. K. Evers, Department of Marketing, Haas School of Business, University of California, Berkeley, 2220 Piedmont Avenue, Berkeley, CA 94720, United States. Email: [evers@haas.berkeley.edu](mailto:evers@haas.berkeley.edu)

when equivalent outcomes can be expressed in different units that make them seem more or less equal. Using these situations of interchangeability allows us to test whether fairness judgments are coherent—in the sense that the same inputs lead to the same outputs across judgments—or whether judgments are influenced by surface characteristics that may lead to inconsistent judgments. Note that our predictions are agnostic toward the underlying moral framework people operate in (e.g., equity vs. equality) and concern themselves with the degree to which judgments based on these frameworks are consistent or inconsistent across different units of expression.<sup>1</sup>

Although situations of interchangeability might seem to be uncommon, we argue they occur more frequently than one might expect. Often, policymakers and decision-makers only select one possible unit of expression when allocating resources, which results in people only encountering one unit of expression from a multitude of options. For example, companies can choose to express bonuses (and punishments) in monetary values (i.e., a simple lump sum of money), paid time off, or additional bonus month(s) of income.<sup>2</sup> In the judicial system, defendants are punished for illegal behavior with direct monetary fines, or time-based punishments (e.g., prison time, day-fines, or a number of months of profit in the case of companies). And service providers can decide whether to compensate failures with a specific cash value or a proportion of the service price.<sup>3</sup> In each of these cases, and more, outcomes distributed equally in one unit often necessarily entail an unequal split in the other. If people use allocative equality to judge fairness (Messick, 1993), then could the same bonuses and fines appear more or less fair depending on how they are expressed? In other words, are peoples' fairness judgments *coherent*<sup>4</sup> across different units of expression?

Although we are, to our best knowledge, the first to investigate how the interchangeability of units affects fairness judgments specifically, a review of related literatures does not lead to a straightforward prediction. On one hand, it is known that when people make decisions that are important to them, they typically engage in System 2 processing, resulting in more coherent decisions (Chaiken, 1980; Shah & Oppenheimer, 2008). Because moral preferences are often more strongly held than more mundane preferences (e.g., Skitka, 2010) this would imply a type of careful reasoning that leads judgments to be unaffected by arbitrary characteristics of the decision-making context such as units of expression. On the other hand, there is evidence that despite the importance people place on moral beliefs, their moral attitudes do not always follow from careful reasoning, but rather are based on intuitive responses that are later rationalized (e.g., Haidt, 2001). If this latter perspective is true, this would imply that, despite its importance, judgments about fairness could be strongly affected by irrelevant cues, such as the unit of expression, and as a result be quite unstable.

Here, we explore how people judge the fairness of distributions and how this affects our understanding of fairness preferences. Because we are interested in evaluations of the fairness of distributions of outcomes, we focus on what is commonly known as “outcome fairness” rather than distributive fairness (for a review, see Skitka et al., 2003), but of course we are not claiming that outcomes are the *only* consideration in judging what is or isn't fair.

We tested the influence of units of expression on fairness judgments in five sets of studies. In Studies 1a and 1b, we elicited

fairness judgments across a variety of contexts and found that participants predominantly judged fairness based on perceived equality in the allocation of expressed units. In Study 2, we built on Study 1 by explicitly converting the salient outcome from one unit to the other (e.g., points to money or money to points). We found that although this did attenuate the effect of units of expression on judgments of fairness, there is a substantial remaining preference for surface-level equality (“arbitrary fairness”). In Studies 3a and 3b, we tested whether the preference for arbitrary fairness extended to distributing outcomes (rather than merely evaluating them) and found that it did. Study 4 is an incentive compatible study showing that people awarded themselves cash bonuses differently depending on the salience of units of allocation. Finally, in Study 5, we tested the implications of these results in the context of policy acceptance. We found that participants were more supportive of income-based fines, when these fines are presented as representing the same cost in time to each perpetrator, rather than different dollar amounts based on their income.

We report all conditions and all variables measured. We predetermined the sample size for all studies and collected data until reaching the predetermined number. All studies except for Study 1a and 3a were preregistered. Because we are studying a novel effect, but expected a relatively large effect size, we opted to typically preregister collecting 100 participants per cell in studies testing for a main effect, and 200 per cell in studies testing interaction effects. Note that because of the way MTurk approves workers, we usually have slightly more participants than the number requested. For studies that included attention checks, we report the preregistered analyses first and where relevant report additional analyses with different exclusion criteria in the main body of the paper or in the online supplemental materials. These studies were approved by the University of California, Berkeley IRB #2018-09-11412. All data and materials are available at: <https://osf.io/9tcbx/>.

## Study 1a

First, we tested the preference for arbitrary fairness based on the allocative unit of expression in the context of punishments.

<sup>1</sup> Because equal outcomes are most commonly documented as representing what people believe to be fair, we expected our participants to typically behave in a similar fashion. Our hypothesis as well as the results in our studies are robust to alternative beliefs about fair distributions at baseline. In the limitations to generalizability section in the general discussion we discuss how arbitrary fairness would express itself in cases where other distributions are believed to be fair.

<sup>2</sup> For example, in The Netherlands, employees typically receive a bonus month's pay, whereas in the United States monetary bonuses appear to be more commonly distributed in set numerical amounts.

<sup>3</sup> For example, in the EU a delayed flight entitles you to a minimum €250 compensation whereas a delayed cruise entitles you to a minimum compensation of 25% of the ticket price.

<sup>4</sup> Judgments are coherent when they are internally consistent (e.g., Dawson & Gregory, 2009); when the same input leads to the same output. This is different from more typically studied biases in moral judgment, such as the self-serving bias (Folger et al., 1983) in which people, consistently, believe allocations to be more fair when it directly benefits them. For a discussion, see Dawson and Gregory (2009).

## Method

In Study 1a, we recruited 201 participants (117 male, 83 female,  $M_{\text{age}} = 36.7$ ,  $SD = 11.3$ ) on Amazon Mechanical Turk (MTurk) and randomly assigned them to one of two conditions. In both conditions we presented participants with a scenario describing two companies, Company A and Company B. Participants read that Company A and Company B are chain stores in the Netherlands that have recently been caught intentionally underpaying their taxes. They were also informed that Company A operates in 28 locations and operates with a daily profit of €1,500 and Company B operates in 24 locations and operates with a daily profit of €1,000. Next, participants read that the government of the Netherlands decides to fine these companies for their tax evasion. In the *money* condition participants read that “Company A was fined €75,000 and Company B was fined €55,000.” In the *time* condition, participants read that “Company A was fined 50 days’ worth of profits and Company B was fined 55 days’ worth of profits.” Importantly, a €75,000 fine for Company A is equivalent to 50 days’ worth of profit. Similarly, for Company B, a €55,000 fine is equivalent to 55 days’ worth of profits. That is, the fines were equivalent between conditions and only differed in their unit of expression. Participants evaluated the fairness of the fines on a 101-point scale ranging from  $-50$  “Extremely unfair to Company A” to  $+50$  “Extremely unfair to Company B” (the zero point was labeled as “Fair to both”). Finally, to ensure that any pattern of responding was not due to participant inattention, participants completed both a memory check and an attention check. The memory check required participants to remember a detail from the scenario (That Company A had a higher daily profit than Company B) and the attention check asked participants to indicate how frequently they have had a fatal heart attack (“never” was the only acceptable response).

## Results

Although this study was not preregistered, we included both a memory check and an attention check item in the survey. We excluded data from participants who incorrectly responded to either of these two items. Applying these exclusions yielded a final sample size of 171 participants.<sup>5</sup> As can be seen in Figure 1, when the punishment was administered in money, participants believed the fine to be unfair to Company A (which had to pay €75,000 as compared with Company B’s €55,000;  $M = -4.9$ ,  $SD = 18.6$ ). However, when the fine was levied in days’ worth of profits, participants believed the fine to be unfair to Company B (which was fined 55 days’ worth of profits compared with Company A’s 50 days’ worth of profits ( $M = 17.4$ ,  $SD = 18.5$ ),  $t(169) = 7.89$ ,  $p < .001$ ,  $d = 1.21$ ,  $\text{diff} = 22.4$ , 95%  $\text{CI}_{\text{diff}} [16.8, 28.0]$ ). In other words, the exact same outcomes were evaluated differently depending on which company *appeared* to be fined more harshly in the units the punishment was expressed in.

### Study 1b

One possible explanation for the results of Study 1a could be that people who are unsure about what the fair punishment ought to be assume that governments and policymakers have already determined for which violations the punishment should be expressed in

time, and for which it should be expressed in money. If this is what our participants believed, then the judgments described in Study 1a would not be incoherent because the units used are not truly perceived to be interchangeable.

To test this possibility, we conducted a study in which we used a clearly arbitrary unit: points (see also Furlong & Opfer, 2009; Hsee et al., 2003). Additionally, this study examined judgments over gains rather than losses.

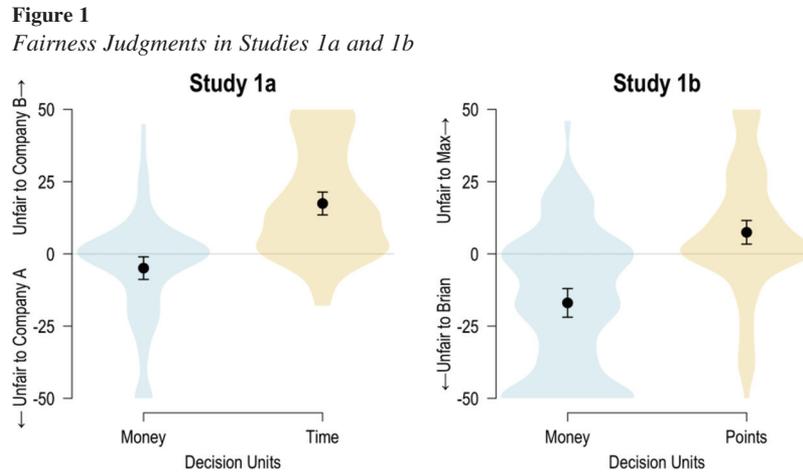
## Method

We recruited 200 MTurk workers (97 male, 103 female,  $M_{\text{age}} = 35.2$ ,  $SD = 11.6$ ) and randomly assigned them to one of two conditions. All participants read about three students who participated in a psychology experiment and who were randomly assigned to one of three roles. Two of the students were randomly assigned to the role of worker 1 (Brian) and worker 2 (Max). Worker 1 would earn \$0.50 per point, whereas worker 2 earned \$1 per point. The third student, Don, was a supervisor who, depending on the condition, was described as being asked to assign a \$27.50 or 40-point bonus to the two students if they each completed at least 20 tasks. In the *money* condition, participants read that Brian received \$12.50 while Max received \$15. In the *points* condition, they read that Brian received 25 points while Max received 15 points. Because these points have no meaning other than representing value within this experiment, the two situations are logically equivalent. All participants then judged the fairness of the allocation on a 101-point scale anchored on  $-50$ : *extremely unfair to Brian*, and  $+50$ : *extremely unfair to Max*.

## Results and Discussion

Consistent with Study 1a, participants judged the bonus to be unfair to Brian in the money condition ( $M = -16.96$ ,  $SD = 24.1$ ) but unfair to Max in the points condition ( $M = 7.45$ ,  $SD = 24.1$ ),  $t(198) = 7.52$ ,  $p < .001$ ,  $d = 1.07$ ,  $\text{diff} = 24.4$ , 95%  $\text{CI}_{\text{diff}} [18.0, 30.8]$  (see Figure 1). Similar to Study 1a, participants judged fairness by relying on the salient unit in which the outcome was expressed. Three additional studies in the online supplemental materials tested the generalizability of the effect in different contexts. Taken together, these studies provide evidence that the exact same outcomes can be judged to be both fair and unfair depending on the units they are expressed in. Of course, this leaves questions as to why this is the case. One possibility is that people are simply confused about the interchangeability of units. This would be something akin to the “miles per gallon illusion” (Larrick & Soll, 2008), which shows that people are quite bad at understanding miles per gallon information, but that recalculating miles per gallon to gallons per mile improves decisions. In the miles per gallon illusion, people made judgments based on salient units rather than their underlying meaning (Kahneman & Frederick, 2002) but more careful deliberation on the relationship between those units resulted in better decisions. It is possible that our participants in Study 1 similarly were motivated to correctly use unit-based information, but lacked the

<sup>5</sup> If we do not apply these exclusions the results are similar. Participants rated the money fine as unfair to company A:  $-4.34$ ,  $SD = 19.6$ , and the (equivalent) time fine as unfair to company B:  $17.9$ ,  $SD = 18.5$ ,  $t(199) = 8.27$ ,  $p < .001$ .



*Note.* The left side of the figure presents the means and standard errors for fairness judgments in Study 1a, rated on a 101-point scale. The right side of the figure presents the distributions of participants' fairness ratings in Study 1b. Fairness judgments differed based on the unit of expression. The black dot in each plot represents the mean, and the whiskers the standard error of the mean. See the online article for the color version of this figure.

ability to take this information correctly into account (e.g., Petty & Cacioppo, 1986). If the results of Study 1 are a consequence of this kind of misunderstanding, then explicitly translating outcomes between units should remove any incoherence between units of expression. This is what we test next.

## Study 2

In Study 2 we tested whether being more transparent about the relationship between the different units of expression and explicitly providing “translations” from one unit into the other would make judgments coherent.

## Method

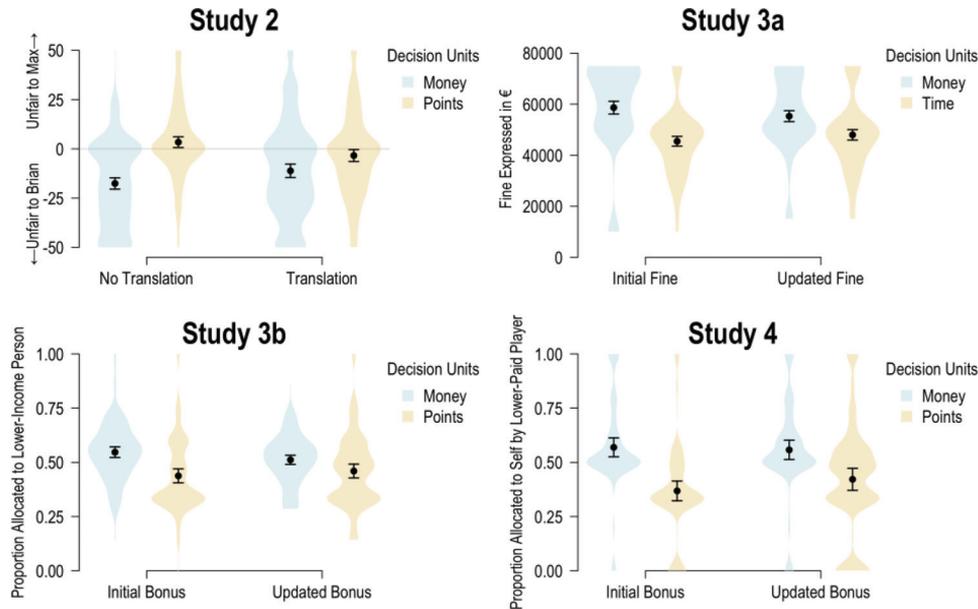
We recruited 813 workers from MTurk (394 male, 416 female, two nonbinary, one missing;  $M_{\text{age}} = 35.9$ ,  $SD = 11.6$ ) and randomly assigned them to one of four conditions. This study used a 2 (units: points or money) by 2 (translation: provided or not) fully between-subjects design. In all conditions, participants read the same scenario as in Study 1b in which Brian earned \$0.50 per point while Max earned \$1.00 per point. Like in Study 1b, participants in the *money* conditions read that Brian received \$12.50 while Max received \$15, whereas participants in the *points* conditions read that Brian received 25 points while Max received 15 points. Participants in the two control conditions indicated how fair they thought the bonus was when expressed in points (in the points condition) or money (in the money condition). In the translation conditions, participants saw the same bonus, but before assessing its fairness we explicitly translated the bonus into the other unit. For example, the participants who saw the 25-point bonus to Brian and 15-point bonus to Max were told that because Brian gets \$0.50 per point and Max \$1 per point, this results in a \$12.50 bonus for Brian, and a \$15 bonus for Max. All participants then rated the fairness of the bonus.

## Results

The results of Study 2 are presented in Figure 2, top left panel. We submitted participants' fairness ratings to an ANOVA, which revealed a main effect of the unit of expression,  $F(1, 809) = 87.0$ ,  $p < .001$ ,  $\eta_p^2 = .10$ , no main effect of adding the translation,  $F(1, 809) = 0.01$ ,  $p = .91$ , and a significant interaction effect,  $F(1, 809) = 18.41$ ,  $p < .001$ ,  $\eta_p^2 = .02$ . An inspection of the means reveals that, replicating the results of Study 1b, participants judged the bonus expressed in points to be unfair to Max ( $M = 3.4$ ,  $SD = 19.3$ ), whereas the same bonus expressed in money was judged as unfair to Brian instead ( $M = -17.6$ ,  $SD = 21.6$ ),  $t(408) = 10.29$ ,  $p < .001$ ,  $d = 1.02$ ,  $\text{diff} = 21.0$ , 95%  $\text{CI}_{\text{diff}} [17.0, 25.0]$ . We next turn to the two “translation” conditions in which the bonus was again expressed either in points or in money, but crucially, where for each bonus we also provided a “translation” into the other unit. In other words, in both conditions, participants saw the bonus expressed both in money and in points, with the only difference being the unit in which the outcome was initially expressed. When the bonuses were instead presented first in one unit and then translated into the other unit, participants judged both to be unfair to Brian, but less so than when the bonus was expressed in points ( $M = -3.4$ ,  $SD = 21.9$ ), as compared with in money ( $M = -11.1$ ,  $SD = 24.5$ ),  $t(401) = 3.35$ ,  $p < .001$ ,  $d = 0.33$ ,  $\text{diff} = 7.8$ , 95%  $\text{CI}_{\text{diff}} [3.2, 12.3]$ . In other words, even when the relationship between units was made completely transparent and participants had full (and identical) information, judgments were still influenced by the unit of expression, making these results hard to reconcile with models that assume insufficient updating (e.g., Van den Bos, 2001) or differences in inferences based on the presented unit. In addition to our hypothesized findings, there also appears to be a general main effect such that participants believe the situation is generally less fair to the lower-income worker. We believe this is probably attributable to the higher-income worker being treated better than the lower-income worker even before the bonus is assigned.

**Figure 2**

*Judgments of Fairness and Outcomes Assigned Differed Based on the Unit of Expression, Even After Participants Saw Their Judgments Translated in the Alternative Unit and Had a Chance to Update Their Decisions*



*Note.* The black dot represents the mean, and the whiskers represent one standard error of the mean. See the online article for the color version of this figure.

Overall, the results of Study 2 suggest that when judging the fairness of allocations, these judgments do not follow from a coherent set of beliefs about what is fair but appear to be affected by what *should* be irrelevant information such as the surface equality of outcomes. A set of conceptual replications of Study 2 using a within-subjects design and using fines versus community service as the units are included in the online supplemental materials.

### Study 3

In Studies 1 and 2, participants rated the fairness of outcomes. We next tested whether the preference for arbitrary fairness would emerge when participants assigned outcomes rather than merely evaluated them. First, if the inconsistencies documented in Studies 1 and 2 also emerge in situations in which decision makers assign outcomes, this has strong practical implications because it suggests that actual policies may similarly be affected by unit of expression. In addition, evaluating outcomes could be expected to involve less cognitive effort than actively assigning outcomes. In that sense, Study 3 can be seen as a test of generality.

Moving from the evaluation of outcomes to the active assigning of outcomes does bring with it a methodological complication, as our question of interest concerns how outcomes are evaluated relative to each other. For instance, suppose a manager gives two employees a month's wages worth of a bonus. If employee A earns \$4,000 a month and employee B earns \$7,000, we predict that translating from "a month's bonus each" (i.e., time) to the respective dollar amount will result in the manager reducing the monetary difference in compensation between employee A and

employee B. There are several ways in which the manager could accomplish this. The manager could increase employee A's bonus while keeping employee B's bonus the same; increase both bonuses but enhance employee A's bonus more than employee B's; or reduce the bonus for employee B. In each of these cases, the goal is not merely to grow the overall bonus, but to increase the relative share of the bonus allocated to employee A. In other words, in the case of assigning outcomes, we would expect that decision makers will try to reduce the perceived difference in outcomes between the targets. To limit the number of possible remedies for (perceived) unfair allocations, we first ran a study in which participants were constrained such that they only assigned the outcome to one target while the outcome for the other target was fixed (Study 3a). We then ran an experiment in which participants assigned outcomes to both targets (Study 3b), which allowed us to study how participants change the relative allocation between the two targets.

### Study 3a

#### Method

We recruited 401 workers from MTurk (184 female, 214 male, three nonbinary,  $M_{\text{age}} = 40.3$ ,  $SD = 12.3$ ). We assigned participants to one of two conditions in a 2 (units—between) by 2 (time—within) mixed design. Participants read about C&A and Kaza, two chain stores they were told were caught participating in tax evasion. In both conditions, participants read that C&A has a daily profit of €1,500 while Kaza has a daily profit of €1,000. Participants in the

money condition then saw that the government decided to fine C&A €75,000 and were asked to assign a fine to Kaza using a free response text box. Participants in the time condition read that the government decided to fine C&A 50 days of profit and were subsequently asked to fine Kaza a number of days of profit using a free response text box. After assigning the fine, all participants saw both the given fine for C&A and their assigned fine for Kaza translated into the other unit. For example, a participant who decided to fine Kaza €60,000 saw that the given €75,000 fine for C&A meant that the company would essentially forego 50 days of profit, and the assigned €60,000 fine for Kaza would mean that Kaza would forego 60 days of profit. After seeing this information, participants again assigned a fine in the same unit as their initially assigned fine.<sup>6</sup> If fairness perceptions are influenced by the unit in which outcomes are expressed, we would expect participants to give Kaza a higher fine in the money condition as compared with the time condition. In addition, as in Study 2, we expected an explicit translation between units to attenuate this effect, but not fully eliminate it.

## Results

On inspection of the data, it was apparent that, because participants were not bound in their responses, some respondents provided values that were extreme outliers.<sup>7</sup> We therefore opted to Winsorize the data at the 5th and 95th percentiles, although we had not preregistered this decision. Specifically, we converted the fines in time into fines in money by multiplying these values by €1,000. Then, we looked at the overall distribution of monetary fines for both conditions (fines assigned in money and fines assigned in time then converted to money). For time 1, we recoded all values less than €10,000 as €10,000 and all values greater than €75,000 as €75,000. For time 2, we recoded all values less than €15,000 as €15,000 and all values greater than €75,000 as €75,000. We conducted, and report in the online supplemental materials, an alternate analysis in which we exclude these outliers rather than Winsorize them and replicate the same effects with this alternative exclusion criterion. We then conducted a repeated-measures analysis using the Mixed procedure in Stata, which used a Satterthwaite approximation to generate degrees of freedom. This analysis predicted participants' assigned fines from their unit condition, the time the fine was assigned (i.e., time 1 and time 2) and an interaction term between these factors. As expected, participants assigned and updated their fines differently, depending on the salient unit of expression (see Figure 2, top right panel). First, the fines were lower for Kaza when participants fined the company in days of profits ( $M = €45,463$ , representing 45.5 missed days of profit,  $SD = €13,965$ ) rather than in Euros ( $M = €58,646$ ;  $SD = €17,902$ ),  $t(588.43) = 8.48$ ,  $p < .001$ ,  $d = 0.82$ , 95%  $CI_{Diff}$  [10,129.24, 16237.58]. As expected, this difference was reduced at time 2 (when participants were presented their judgments translated into the other unit), with participants in the time condition increasing the fine to an average of €47,980, representing 48 missed days of profit ( $SD = €14,874$ ) and participants in the money condition decreasing the fine to an average of €55,303 ( $SD = 15,315$ ), representing 55 days of missed profit. This attenuation was confirmed by a significant interaction effect;  $t(588.43) = 5.44$ ,  $p < .001$ . Although the difference in fines attenuated, participants still assigned significantly different fines at

time 2;  $t(588.43) = 4.71$ ,  $p < .001$ ,  $d = 0.49$ , 95%  $CI_{Diff}$  [4,268.561, 10,376.91].

For exploratory reasons, we also included a question at the end of the survey which asked participants how they believed fines *should* be imposed in situations like these with the following three response options: similar crimes should be fined the same monetary value; similar crimes should be fined equal lengths of time; and profit should be taken into account even if that leads to different amounts of fines or length of time. Most participants in both conditions (61%) indicated that they believed profit should be taken into account, even if that leads to dissimilar fines. If we restrict our analyses to only those participants who indicated that policymakers should actively take profit into account when assigning fines, we still find the same effects. This subset of participants still assigned different fines depending on the salient unit, and partially attenuated this difference after the fine was translated into the other unit of expression.

## Study 3b

In Study 3a, we found that participants' allocations of fines were influenced by the unit of expression. Providing participants with the fines expressed in the alternative unit attenuated but did not eliminate the reliance on salient unit. Because decision-makers have multiple ways to equalize outcomes, Study 3a restricted participants judgments to one target, which makes the results easier to interpret within the experiment, but limits generalizability. In Study 3b participants assigned outcomes to both targets, which allowed us to see how participants change the outcomes for each target relative to the other.

## Method

We recruited 229 workers from MTurk (91 female, 137 male, one missing,  $M_{age} = 33.5$ ,  $SD = 10.7$ ). We used the same scenario in 3b as in Study 1b and Study 2, but rather than evaluating the bonuses assigned to Brian and Max, participants in this study assigned the bonuses themselves. Specifically, participants were either assigned to a points condition, in which they were asked to distribute 40 points between Brian and Max, or a dollar condition in which they were asked to distribute \$35 between Brian and Max. Just like in Studies 1b and 2, points were worth \$0.50 to Brian and \$1.00 to Max. As in Study 3a, after assigning the bonuses, we calculated the translations of the bonus into the other unit of expression and asked participants to assign the bonus once more, allowing us to test whether explicitly spelling out the translation of the initial assignment would reduce the reliance on surface equality.

<sup>6</sup> In the majority of the studies reported here we told participants it was completely fine to give the same answer as they gave before, because we were worried asking participants to again indicate a distribution would lead to demand effects to update. A reviewer pointed out that inclusion of this sentence could lead to opposite demand effects, so this study did not include that sentence.

<sup>7</sup> The full, original dataset is available on the OSF project page for this article; inspection of these outliers suggests participants accidentally answered the money question with a fine in days and vice versa, because these outliers were typically off by a factor of 1000. These participants also showed other "red flags" such as completing the survey through a proxy and broken English on an open-ended language check.

## Results

The results of Study 3b are similar to those of Study 3a (see Figure 2, bottom left panel). Because we are interested in the relative share of the overall bonus distributed to each worker and the total amount of money distributed varies across participants, we analyzed the proportion of the total amount of money assigned to each worker. For example, if a participant assigned \$20 to Brian and \$15 to Max, we would code this as 57% of the bonus being assigned to Brian (and thus 43% to Max). Similarly, if someone assigned 30 points to Brian (equaling \$15) and 10 points to Max (equaling \$10), they would be coded as assigning 60% of the total bonus to Brian (and thus 40% to Max).<sup>8</sup>

If the unit of expression does not affect judgments, we would expect participants to assign the same proportion of the bonus to Brian, regardless of the unit of expression. Instead, we found that at time 1, Brian (the lower-income person) was assigned a much lower proportion of the bonus in the points condition (43.8%,  $SD = 17.4\%$ ) than in the money condition (54.7%,  $SD = 13.5\%$ ),  $t(227) = 5.48$ ,  $p < .001$ ,  $d = 0.65$ ,  $diff = 10.9\%$ , 95%  $CI_{diff}$  [7.0%, 14.8%]. As in previous studies, this difference was attenuated at time 2, but Brian still received a lower bonus in the points condition (46.0%,  $SD = 17.3\%$ ) than in the money condition (51.2%,  $SD = 11.6\%$ ),  $t(227) = 2.60$ ,  $p = .01$ ,  $d = 0.36$ ,  $diff = 5.2\%$ , 95%  $CI_{diff}$  [1.4%, 9.0%]. This pattern was qualified by a significant interaction effect;  $F(1, 227) = 11.71$ ,  $p < .001$ ,  $\eta_p^2 = .05$ .<sup>9</sup> As in the previous studies, when we provided participants with a translation of their assigned outcomes from one unit into another unit, participants reduced the difference in assigned bonuses, but did not fully eliminate them.

## Discussion

In Studies 3a and 3b (and Study S3A in the online supplemental materials), we find that participants assign starkly different fines and bonuses depending on the salient unit of expression. Although explicitly translating the outcomes between units of expression, and thus showing the consequences of their decisions on the non-salient unit, did reduce these differences somewhat, participants with identical information still assigned different bonuses and fines dependent on the unit of expression. For example, in Study 3a, the punishment for the tax-evading company was about 15% higher when the company was fined in money rather than days of profit. This persistent difference reflects a repeated preference of participants to equalize punishment on the salient unit, rather than the underlying relationship between the units.

### Study 4

So far, we have found that when evaluating and when distributing outcomes, people appear to rely on perceived surface equality. In Study 4 we tested whether a preference for arbitrary fairness would emerge in an incentive-compatible setting, in which participants had financial stakes in the outcome.

## Method

### Participants and Design

Owing to the high cost of this study, we preregistered a sequential analysis (Lakens, 2014). Under this plan we would analyze the

data at predetermined stopping points, with a lower  $p$ -value threshold for significance at each sequential analysis. We ultimately approached 200 pairs of participants on a large university campus and invited them to play a game of “cornhole” in which each participant made 16 attempts each to throw a small bag of corn through a hole in a wooden board. We were interested in how participants would distribute a bonus earned by the pair of them. For each pair of participants, only one person distributed the bonus. These distributors were 130 males, 69 females, one missing;  $M_{age} = 22.8$ ,  $SD = 6.6$ . We used a 2 (units—between) by 2 (time—within) mixed design.

### Procedure

Participants played collaboratively in teams of two and were told that they would earn money per point scored. They were also informed that if each of them scored at least three points, they would win an additional bonus. We then told them that there was a twist: one of them would earn \$0.25 per point scored, and the other \$0.50 per point. We flipped a coin to randomly and transparently assign participants to roles. Nearly all groups earned the bonus. Then, the \$0.25 per-point player was taken aside and told they could divide a bonus of \$5 (in the *money* condition) or 14 points (in the *points* condition) between themselves and the other team member. Importantly, if the \$0.25 per point player relied on superficial equality in allocating the points-based bonus, this would work against their interests and result in a lower bonus for themselves. As in Study 3b, after participants assigned the bonus, the research assistants translated the assigned outcomes to the other unit and provided participants with the opportunity to change their allocations (although, of course, they did not have to do so).

## Results

The results of this incentive-compatible study were very similar to the results of Study 3b. Participants in the points condition initially gave themselves a smaller share of the bonus (36.9%) compared with participants in the money condition (56.9%),  $t(197) = 6.07$ ,  $p < .001$ ,  $d = 0.88$ ,  $diff = 20.1\%$ , 95%  $CI_{diff}$  [13.7%, 26.4%]. This difference between conditions became smaller at time 2, interaction  $F(1, 197) = 9.63$ ,  $p = .002$ ,  $\eta_p^2 = .05$ , but even at time 2 the allocators still gave themselves a smaller share of the bonus in the points condition (42.2%) compared with the money condition (55.8%),  $t(197)^{10} = 4.13$ ,  $p < .001$ ,  $d = 0.56$ ,  $diff = 13.6\%$ , 95%  $CI_{diff}$  [6.8%, 20.4%]. In fact, even after seeing the explicit

<sup>8</sup> We opted to code outcomes as the proportion of the total dollar value received rather than points because we believe the proportion of money received is ultimately the more important outcome. However, because our predictions all involve relative differences, recoding the data to proportion of points received rather than the proportion of money received shows the same effects, due to the inherent relationship between the two units.

<sup>9</sup> Here we present the data as the proportion of the total monetary bonus assigned; we could of course present it as the proportion of total points assigned instead which would change the means, but not the crucial relative effect of interest: that judgments are more inconsistent at Time 1 vs. Time 2, and that despite this attenuation there still is an effect of condition at Time 2.

<sup>10</sup> For one participant, the translation from points to money was miscalculated by the research assistants, so this participant was dropped from the analysis.

translation, participants in the points condition still gave themselves less than half of the bonus in dollar terms.

Because participants in this study were recipients of the bonus (rather than third-party observers), they could try to maximize shared outcomes (i.e., increase the overall size of the pie) by assigning as many points as possible to the other (higher income) player. Although some participants did this, if the initial difference in outcomes between the two conditions was because participants were motivated to maximize absolute rather than relative outcomes, then this difference should have *increased* (or stayed the same) when participants were provided with more information, rather than decreased. In other words, participants attempting to maximize combined outcomes (e.g., Charness & Rabin, 2002; Fehr & Schmidt, 1999) would work against our predicted pattern of results and thus make this design a conservative estimate of the reliance on salient units. Overall, Study 4 provides strong evidence for the arbitrary fairness effect, finding evidence for the influence of unit of expression even in a situation where the underlying relation and its implications are fully transparent, and where reliance on surface equality directly reduces the participants outcomes.

### Study 5

So far, we have investigated how people judge different allocations and how they allocate outcomes themselves under conditions of interchangeability. One implication of our findings is that the exact same schemes of reward and punishments could be made to be judged as more or less fair by drawing attention to units in which these outcomes appear more equal. For example, in the case of legal violations, sentences are often equalized on time (e.g., mandatory minimum prison sentences require the same time served for all, regardless of income) or on money (e.g., traffic fines cost the same amount of money for all violators regardless of income). One possible implication of the preference for arbitrary fairness is that income-based fines might become more palatable when they are reframed as equal time of work rather than unequal amounts of money. We tested these implications for policymakers in Study 5.

### Method

We recruited 311 workers on MTurk (179 males, 129 females, three nonbinary;  $M_{\text{age}} = 35.1$ ,<sup>11</sup>  $SD = 11.2$ ) and randomly assigned them to one of three conditions. In all conditions, participants read about a policy the government of Luxembourg was considering imposing to reduce reckless driving. In the time condition, participants read that the government was considering sentencing those who engaged in extreme speeding on the freeway to 6 days in jail. Next, we explained that a person who earns €2000 per month would be sent to jail for 6 days, and a person who earns €4,000 per month would also be sent to jail for 6 days. This condition served as a baseline in which outcomes were equal on the salient unit of expression. In the money condition, participants read that the government was considering imposing a fine on those caught speeding of 20% of their monthly income. Then, we explained that this meant that a person who earns €2,000 per month would have to pay a €400 fine, whereas a person who earns €4,000 per month would pay a €800 fine. In the final condition, the time-to-money condition, participants read that the government planned to send speeders to jail for 6 days. However, we explained to participants

that to reduce costs, the government instead opted to replace the jail time with a fine that would resemble the money the person would lose out because of missed wages by being in jail. Then, we explained that because 6 days constitutes about 20% of the month, the government instead decided to fine extreme speeders 20% of their monthly income. For a person who earns €2,000 per month, this would constitute a €400 fine, for a person who earns €4,000 per month, this would constitute an €800 fine. Finally, after reading the policy proposal associated with their assigned condition, all participants indicated the degree to which they felt the policy was acceptable on a 101-point scale anchored at  $-50$  *completely unacceptable*, and  $+50$  *extremely acceptable*.

### Results

As expected, participants in the money condition perceived the policy to be less acceptable ( $M = 1.1$ ,  $SD = 33.9$ ) compared with participants in the time condition ( $M = 10.3$ ,  $SD = 32.3$ ),  $t(308) = 2.01$ ,  $p = .045$ ,  $d = 0.28$ , 95%  $CI_{\text{diff}}$  [.005, 18.43]. Crucially, participants in the money condition also perceived the policy to be less acceptable ( $M = 1.1$ ,  $SD = 33.9$ ) compared with participants in the time-to-money condition ( $M = 10.2$ ,  $SD = 31.2$ ),  $t(308) = 2.04$ ,  $p = .043$ ,  $d = 0.28$ , 95%  $CI_{\text{diff}}$  [.27, 17.98]. The time condition did not differ significantly from the time-to-money condition;  $t(308) = 0.02$ ,  $p = .98$ ,  $d < .01$ . Note that despite only varying the units in which the outcomes are expressed and not providing participants with any actual information about the procedures resulting in those outcomes, it is possible in this study that participants form beliefs about these procedures based on the outcomes (see, e.g., Van den Bos, 2001) and that it is these unit-driven beliefs making the day fine more palatable.

Although we did not preregister a plan to exclude any participants from our analyses, this study included a simple language check.<sup>12</sup> If we restrict our analysis to only include the participants who passed this check (remaining  $N = 283$ ), we do not find meaningfully different results from the preregistered analysis. Specifically; participants in the money condition perceived the policy to be less acceptable ( $M = 0.6$ ,  $SD = 34.3$ ) as compared with participants in the time condition ( $M = 11.1$ ,  $SD = 30.9$ ),  $t(280) = 2.19$ ,  $p = .029$ . Participants in the money condition also perceived the policy to be less acceptable ( $M = 0.6$ ,  $SD = 34.3$ ) as compared with participants in the time-to-money condition ( $M = 10.1$ ,  $SD = 31.9$ ),  $t(280) = 2.01$ ,  $p = .045$ . The time condition did not differ significantly from the time-to-money condition;  $t(280) = 0.21$ ,  $p = .83$ .

### General Discussion

In seven experiments (and an additional seven reported in the online supplemental materials), people showed a preference for equating outcomes that could be expressed in multiple units on whichever unit is salient, a preference we call *arbitrary fairness*. People showed this preference when evaluating the fairness of others' allocations (Studies 1a, 1b, and 2) and when allocating

<sup>11</sup> One person indicated their age as 1982. Because this study was conducted in 2018, we coded this person as being 36 years old.

<sup>12</sup> Please describe three things you see in this picture with a picture of a beach scene.

outcomes themselves (Studies 3a, 3b, and 4). This preference persisted even when the units were arbitrary (e.g., points), when outcomes were translated between units for participants, and when equal allocation of resources directly conflicted with the participants' self-interest. The preference for arbitrary fairness thus appears to be a general and robust phenomenon. An important implication is that the same punishment can be deemed more or less acceptable depending on how this punishment is presented (Study 5).

### Integration With Existing Research

A fundamental question in judgment and decision-making research is what are preferences and how are they formed. The dominant perspective is that human decision-making typically follows one of two processes. The first, often called System 1 or the intuitive system, is characterized by fast decisions, often based on intuition. System 2, on the other hand, is characterized by slower, deliberate processing. When decisions are not important, we typically use System 1 processing, saving cognitive resources in the process. Commonly, violations of coherence are the result of System 1 processing, and these biases in decision-making disappear when people deliberate more and switch to System 2 processing instead (See e.g., Kahneman & Frederick, 2002; Stanovich & West, 2000). For example, people think that increasing the cost of gasoline by \$0.10 to reduce deaths from air pollution is more justified when casualties will be reduced from 20,000 to 10,000 than from 200,000 to 190,000, when they see only one of the two scenarios (Kahneman & Tversky, 2013). Intuitively, going from 20,000 casualties to 10,000 feels like a bigger improvement than going from 200,000 to 190,000. However, when people see both of these scenarios at the same time, they find the price increase to be equally justified in both cases (Frisch, 1993). When it is clear one's intuition leads to inconsistent judgments, people engage in reason-based System 2 judgments. Similarly, when decisions become more important, for example when there is money at stake, people often switch from intuition-based System 1 processing to more deliberate System 2 processing, leading to more coherent judgments (e.g., Camerer & Hogarth, 1999). Finally, sometimes people just lack the ability to engage in System 2 processing. For example, in the miles per gallon illusion, many people do not have the numerical knowledge to make the right judgment, forcing them to rely on simplifying System 1 heuristics (Larrick & Soll, 2008). However, when participants' mathematical ability was facilitated by translation miles per gallon into gallons per mile, allowing participants to use the (correct) System 2 process, judgments improved.

Despite dual process models being the dominant perspective for how people make judgments and decisions, a few phenomena have been documented that do not appear to fit this mold. These are cases in which participants are aware that they are (by normative standards) making an error, but nonetheless do not correct it. This phenomenon has been termed "acquiescence" (Risen, 2016; Walco & Risen, 2017). For example, work on the ratio-bias (Denes-Raj & Epstein, 1994) shows that people often prefer lower probability gambles with more opportunities to win (e.g., 7/100) over higher probability gambles with fewer opportunities to win (e.g., 1/10). Interestingly, some participants explicitly indicated understanding that their chances of winning would be higher in the 1/10 case, but still preferred the 7/100 gamble because the just

"felt" they would have a better chance of winning. Similarly, Walco and Risen (2017) used a Monty Hall paradigm and found that of those who correctly indicated that, based on a rational analysis, switching would increase their chances of winning, 24.5% still decided to stick with their initial decision. We believe that, similarly, the behavior documented in this article reflects judgments being pulled in opposite directions by arbitrary fairness intuitions on the one hand and explicit knowledge about equivalence across units of expression on the other (similar to Criterion "S," Sloman, 1996).

In the studies on acquiescence there is a rationally correct answer; 1/10 is a higher probability than 7/100. In the moral domain there often isn't one correct answer; moral beliefs are inherently subjective. That said, behavior akin to acquiescence has been documented in the moral domain with work on moral dumbfounding (Haidt, 2001) showing a similar unwillingness to deviate from intuition. For example, participants judged a scenario involving a medical student eating a piece of a cadaver she was prepping for an anatomy class. After judging whether the act was moral or not, the authors would question the participants' judgments and "'play devil's advocate', by questioning their reasons" (Haidt et al., 2001, p. 6). They found that this often left participants "dumbfounded," still believing the action was wrong, but acknowledging that they were unable to explain why. One of the participants in our Study 4 volunteered similar information. This participant was the "low-income worker" assigning the bonus between himself and his partner in points, and assigning 7 points each. After we translated this to the respective dollar values and asked him to assign the bonus again, this participant proclaimed; "I realize I am giving myself too little, but it just feels so wrong to not split the bonus evenly!." That said, it has been debated to which degree moral dumbfounding truly is a result of people sticking to their intuitions against their better judgment, or rather a case in which people do not accept the reasons given as valid (Royzman et al., 2015).

We believe our studies circumvent both of these concerns in showing a behavior akin to acquiescence in the moral domain. By using internal consistency (coherence) as a benchmark, we can be agnostic about a correct standard of fairness and instead investigate whether participants' behavior is consistent with their own proclaimed beliefs. In addition, whereas in moral dumbfounding literature the source of conflicting information is external (and can thus be dismissed by the participants), in our studies the source of conflict is self-generated.

At a high level, we believe these results imply that fairness judgments may not be as coherent as previously believed. Most models of fairness assume that fairness judgments arise from some sets of beliefs that are combined in a logically coherent way. For example, in relational models theory (Fiske, 1992) the degree to which people believe an exchange to be fair depends on the relationship between those people. Those in a market pricing relationship believe fair exchanges to involve proportionality whereas those in a communal relationship do not. However, there is an implicit assumption that within these relational models, people are coherent. Likewise, work on procedural and outcome fairness (e.g., Trautmann & van de Kuilen, 2016) distinguishes between inputs to fairness judgments, but again assumes that changes in these inputs leads to changes in judgments in an internally consistent way. Even work investigating sources of bias in fairness judgments still assumes

those judgments to be coherent. For example, looking at procedural justifications, Folger et al. (1983) found that participants in a competition for resources with another participant believed a deviation from procedure to be much more acceptable when it increased their winning chances as compared to when it increased the others' winning chances, suggesting a self-serving bias. Although in this case fairness judgments can be *biased* they are still *coherent*, with outcomes just being judged fairer when they benefit the self. Our article fits within a small set of other work challenging the notion that there may even be such a coherent framework and that it may be more fruitful to investigate seemingly irrelevant features of the decision-making context and how those influence fairness judgments.

### Constraints on Generality

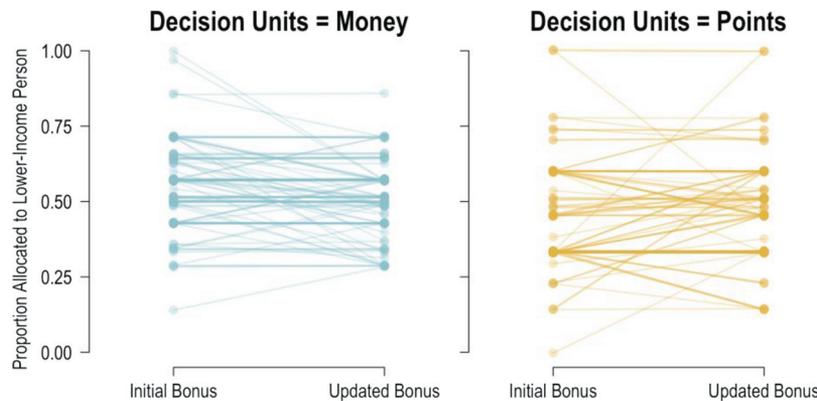
Following Simons et al., 2017, we discuss some possible constraints on the generality of our findings. Although our studies used a variety of different distributions across different contexts involving both losses (punishments) and gains (reward), all our studies used American participants. To which degree can the findings documented be generalized to different cultures? We believe that the core finding in this article—that fairness judgments do not fully reflect the interchangeability of units—to be universal. Although of course this is ultimately an empirical question, the fact that the preference for unit-level equality is so strong that careful deliberation does not eliminate it suggests to us that it is likely present across cultures. That said, we would expect the specific expression of this phenomenon to be affected by culture and context. For example, in most of our studies, participants appear to believe an equal distribution of resources is fair. In cultures with a strong focus on helping those worse off, we would expect different consideration of fairness to affect beliefs about what a fair distribution is (e.g., an unequal distribution in which the person worse off gets a larger share of the pie). However, we would still expect those beliefs to mostly depend on the level of whichever unit in which the distribution is expressed. For example, in such a culture

people might prefer an 80% versus 20% distribution of resources, but we would expect people to prefer this distribution in terms of the more salient unit and, like in the studies reported here, a change in the salience of units to lead to inconsistent judgments. Similarly, we would expect considerable heterogeneity in general support of allocations across domains (e.g., Republicans being less supportive of redistribution as compared with Democrats), but here as well we would not expect the same heterogeneity on the degree to which both Republicans and Democrats rely on salient units of expression, and to be influenced by those salient units to an equal degree.

It is also important to note that we found considerable heterogeneity in how individuals updated their fairness judgments after seeing their judgments translated from one unit into the other. In the between-participants studies we cannot examine updating at the participant level, but in the within-participants studies we can know whether and how participants updated their judgments. First, many participants do not change their judgments at all in response to the translation manipulation (between 36% and 74% across studies do not change their answers between time 1 and time 2). It is not the case, however, that those who do update do so all in the same way (e.g., those in the points conditions all update their judgments to equalize financial outcomes). Instead, we find that the participants who update do so in many different ways: some slightly adjust their initial judgment, whereas others fully equalize on the alternative unit. As an example, see Figure 3 which depicts participants' updating of judgments in Study 3b.

One could also wonder why participants in our studies appear to behave differently from those in Andreoni and Miller (2002). In Andreoni and Miller (2002), participants played dictator games with another person using tokens that reflected different amounts of money. In this work, participants behaved as if they considered the relative value of the tokens, such that dictators gave more tokens when they were worth more to the receiver and fewer when they were worth less. However, we find that participants do not appear to take interchangeability into account, and when they do,

**Figure 3**  
*Participants' Adjusted Distributions From Time 1 to Time 2 for the Low-Income Worker in Study 3b*



*Note.* On the left (in blue) are adjustments in the money condition, on the right (in yellow) are adjustments in the points condition. See the online article for the color version of this figure.

they try to equalize outcomes in expressed units, not maximize overall outcomes. Although our studies differ in several ways that could explain these different findings (e.g., Andreoni & Miller used only economics students, who are known to behave differently from the general population; see, e.g., Frank et al., 1993, 1996), we believe the most important difference between these studies is the degree to which fairness concerns are salient. For example, in our studies involving reward, both actors actively worked together to reach a common goal that resulted in a bonus, whereas in Andreoni and Miller (2002) the receiver in the dictator game was merely an anonymous participant to which people could donate money. In this sense, Andreoni & Miller's work is more of a test of altruism, whereas our work is focused specifically on fairness.

A recent working paper by Exley and Kessler (2018) finds a similar reliance on salient units to the one we describe, albeit using a very different study design. In Exley and Kessler's study, two targets were endowed with a certain amount of "small tokens" worth \$0.01 and "large tokens" worth \$0.02. Using this design, it is possible to have one target possess a larger number of small tokens, but a smaller amount of total wealth, as compared with the other target. Then, they asked participants to remove a certain number of small tokens from the two targets. If participants consider total value, they would be expected to remove tokens from the target with the greatest total wealth. However, what they found was that a sizable proportion of participants instead removed more small tokens from the target holding the largest amount of small tokens, even when that target had lower total wealth. We believe their results are conceptually consistent with our findings and suggest that, when considering fairness, people often only rely on surface equality as a cue without truly considering the implications of this surface equality.

## Policy Implications

We examine preferences for arbitrary fairness in contexts in which the selection of a unit of expression is truly arbitrary, or close to it. However, reducing this constraint only slightly reveals a multitude of situations in daily life where this reliance on surface equality might play a role in forming people's judgments. We believe the most salient domain is the legal system. In legal systems across the world, many types of punishments are either equalized on money when the punishment is in money to be paid (like traffic tickets), or on time when it involves time spent in prison or doing community service (often but not always for more serious crimes). Here we see a conceptually similar (though not perfect) interchangeability. To pay a fine, people have to work for some time to earn money, and likewise the time spent in prison or doing community service is time one could have spent earning money. In the online supplemental materials, we present three additional studies with designs parallel to those of Studies 1–3, and find even stronger reliance on surface equality in comparing fines with community service. Even though community service and fines are not as interchangeable as Euros vs. months of profit (people may naturally consider leisure as an alternative to community service rather than considering that time as a lost opportunity for wages), the emergence of a similar pattern of results suggests that, within the legal domain, the evaluation and assignment of "fair" punishments may be susceptible to surface equality in a

similar way to how participants in our studies were affected by it. This is consistent with other work finding a reliance on heuristic decision-making within the legal domain and suggests our legal system may not act as fair as it intends to (see, e.g., Dhimi & Ayton, 2001).

How *should* punishments and rewards be distributed? Throughout this article, we have discussed situations in which participants' responses are inconsistent with each other, but we have been careful in not specifying which, if either, of their responses is an error. The question of what constitutes a fair reward or punishment is more philosophical in nature than empirical, and to a large degree likely depends on the goal associated with allocating reward and meting out punishments. If the goal is to deter or promote a certain behavior, pain and pleasure should be dispensed equally across people. In those cases, equalizing on experienced impact may be the fair option. However, in other situations the goal may be compensation, in which case equalizing on the cost may be the fair option. For example, when two drivers with a different income both strike a traffic light, if the goal of the punishment is to repair the damage, they should be fined an equal amount of dollars (to restore the damage); yet if the goal of the punishment is to encourage more careful driving, equalizing the fine based on income may be the fairest way to punish (but, although many people explicitly agree with this logic, they often do not act consistently with it; see Carlsmith et al., 2002). Our article does not attempt to determine what is fair. Instead, we identify domains in which decision-makers may not realize that judgments are affected by unit of expression (unfair systems may appear fair and vice versa), while also providing tools on how to make certain reward and punishments more acceptable to the public.

## References

- Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental psychology* (Vol. 2, pp. 267–299). Academic Press.
- Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2), 737–753. <https://doi.org/10.1111/1468-0262.00302>
- Bolton, G. E., Katok, E., & Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, 27(2), 269–299. <https://doi.org/10.1007/s001820050072>
- Brosnan, S. F., & De Waal, F. B. (2003). Monkeys reject unequal pay. *Nature*, 425(6955), 297–299. <https://doi.org/10.1038/nature01963>
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1/3), 7–42. <https://doi.org/10.1023/A:1007850605129>
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284–299. <https://doi.org/10.1037/0022-3514.83.2.284>
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5), 752–766. <https://doi.org/10.1037/0022-3514.39.5.752>
- Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817–869. <https://doi.org/10.1162/003355302760193904>

- Dawson, N. V., & Gregory, F. (2009). Correspondence and coherence in science: A brief historical perspective. *Judgment and Decision Making*, 4, 126–133.
- Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, 66(5), 819–829. <https://doi.org/10.1037/0022-3514.66.5.819>
- Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, 14(2), 141–168. <https://doi.org/10.1002/bdm.371>
- Evers, E., Inbar, Y., & O'Donnell, M. (2022). *Fairness*. Open Science Framework. <https://osf.io/9tcbx/>
- Exley, C. L., & Kessler, J. B. (2018). *Equity concerns are narrowly framed*. Harvard Business School Working Paper, No. 18-040, November 2018. (Revised August 2021)
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868. <https://doi.org/10.1162/003353599556151>
- Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. Free Press.
- Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99(4), 689–723. <https://doi.org/10.1037/0033-295X.99.4.689>
- Folger, R., Rosenfield, D., Rheaume, K., & Martin, C. (1983). Relative deprivation and referent cognitions. *Journal of Experimental Social Psychology*, 19(2), 172–184. [https://doi.org/10.1016/0022-1031\(83\)90036-7](https://doi.org/10.1016/0022-1031(83)90036-7)
- Frank, R. H., Gilovich, T. D., & Regan, D. T. (1996). Do economists make bad citizens? *The Journal of Economic Perspectives*, 10(1), 187–192. <https://doi.org/10.1257/jep.10.1.187>
- Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *The Journal of Economic Perspectives*, 7(2), 159–171. <https://doi.org/10.1257/jep.7.2.159>
- Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, 54(3), 399–429. <https://doi.org/10.1006/obhd.1993.1017>
- Furlong, E. E., & Opfer, J. E. (2009). Cognitive constraints on how economic rewards affect cooperation. *Psychological Science*, 20(1), 11–16. <https://doi.org/10.1111/j.1467-9280.2008.02244.x>
- Haidt, J., Bjorklund, F., & Murphy, S. (2001). *Moral dumbfounding: When intuition finds no reason* [Unpublished manuscript]. University of Virginia.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. <https://doi.org/10.1037/0033-295X.108.4.814>
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). “Economic man” in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815. <https://doi.org/10.1017/S0140525X05000142>
- Hsee, C. K., Yu, F., Zhang, J., & Zhang, Y. (2003). Medium maximization. *The Journal of Consumer Research*, 30(1), 1–14. <https://doi.org/10.1086/374702>
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 49, Article 81.
- Kahneman, D., & Tversky, A. (2013). Choices, values, and frames. In L. C. MacLean, & W. T. Ziemba (Eds.), *Handbook of the fundamentals of financial decision making: Part I* (pp. 269–278). World Scientific. [https://doi.org/10.1142/9789814417358\\_0016](https://doi.org/10.1142/9789814417358_0016)
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710. <https://doi.org/10.1002/ejsp.2023>
- Larrick, R. P., & Soll, J. B. (2008). Economics. The MPG illusion. *Science*, 320(5883), 1593–1594. <https://doi.org/10.1126/science.1154983>
- Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Esminder, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More ‘altruistic’ punishment in larger societies. *Proceedings. Biological Sciences*, 275(1634), 587–590. <https://doi.org/10.1098/rspb.2007.1517>
- Messick, D. M. (1993). Equality as a decision heuristic. In B. A. Mellers & J. Baron (Eds.), *Psychological perspectives on justice: Theory and applications* (pp. 11–31). Cambridge University Press. <https://doi.org/10.1017/CBO9780511552069.003>
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19, 124–129. [https://doi.org/10.1016/S0065-2601\(08\)60214-2](https://doi.org/10.1016/S0065-2601(08)60214-2)
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5), 1281–1302.
- Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, 123(2), 182–207. <https://doi.org/10.1037/rev0000017>
- Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, 10(4), 296–313.
- Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134(2), 207–222. <https://doi.org/10.1037/0033-2909.134.2.207>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, 4(4), 267–281. <https://doi.org/10.1111/j.1751-9004.2010.00254.x>
- Skitka, L. J., Winkler, J., & Hutchinson, S. (2003). Are outcome fairness and outcome favorability distinguishable psychological constructs? A meta-analytic review. *Social Justice Research*, 16(4), 309–341. <https://doi.org/10.1023/A:1026336131206>
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5), 645–665. <https://doi.org/10.1017/S0140525X00003435>
- Trautmann, S. T., & van de Kuilen, G. (2016). Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity. *Journal of Risk and Uncertainty*, 53(2–3), 75–88. <https://doi.org/10.1007/s11166-016-9249-4>
- Van den Bos, K. (2001). Fairness heuristic theory. In S. Gilliland, D. Steiner, & D. Skarlicki (Eds.), *Theoretical and cultural perspectives on organizational justice* (pp. 63–84). IAP.
- Walco, D. K., & Risen, J. L. (2017). The empirical case for acquiescing to intuition. *Psychological Science*, 28(12), 1807–1820. <https://doi.org/10.1177/0956797617723377>
- Walster, E., Walster, G. W., & Berscheid, E. (1978). *Equity: Theory and research*. Allyn & Bacon.

Received June 16, 2021

Revision received August 1, 2022

Accepted August 4, 2022 ■