Draft h1, August 22, 2018

Similarity and the Trustworthiness of Distributive Judgments

Alex Voorhoeve (corresponding author)

Department of Philosophy, Logic and Scientific Method

London School of Economics and Political Science

Houghton Street, London, WC2A 2AE, UK

a [dot] e [dot] Voorhoeve [at-sign] lse [dot] ac [dot] uk

&

Department of Applied Economics and Department of Philosophy

Erasmus University Rotterdam

Arnaldur Stefansson

Department of Economics

Uppsala University

Brian Wallace

Source Clear, San Francisco

Author note: Brian Wallace's work on this paper was carried out while at the Department of Economics, University College London.

Interests: None of the authors have any interests to declare.

Acknowledgements

This paper was presented at Copenhagen University, the Experimental Philosophy Conference at Newcastle University, Fudan University, George Washington University, Harvard University, the Institute of Philosophy in London, LSE, Oxford University, Universidade Nova, the University of Maryland at College Park, Uppsala University, the U.S. National Institutes of Health, and Warwick University. We thank our audiences, Ken Binmore, Luc Bovens, Antonio Cabrales, Ipek Gençsü, Joseph Millum, Erik Schokkaert and two anonymous referees for comments and UCL's Centre for Economic Learning and Social Evolution (ELSE) for the use of its laboratory. We thank Carl Runge for independent coding of subjects' rationales for their choices. This research was supported by the British Academy through grant SG 45949 and by the British Arts and Humanities Research Council through grant AH/J006033/1. Alex Voorhoeve is grateful for a Visiting Scholarship to the Department of Bioethics at the U.S. National Institutes of Health, during which much of this paper was written. The opinions expressed are the view of the authors only. They do not represent any position or policy of the U.S. National Institutes of Health, the Public Health Service, or the Department of Health and Human Services.

Abstract

When people must either save a greater number of people from a smaller harm or a smaller number from a greater harm, do their choices reflect a reasonable moral outlook? We pursue this question with the help of an experiment. In our experiment, two-fifths of subjects employ a similarity heuristic. When alternatives appear dissimilar in terms of the number saved but similar in terms of the magnitude of harm prevented, this heuristic mandates saving the greater number. In our experiment, this leads to choices that are inconsistent with all standard theories of justice. We argue that this demonstrates the untrustworthiness of distributive judgments in cases that elicit similarity-based choice.

Keywords: Similarity, distributive justice, moral decision-making, heuristics, reflective equilibrium.

Introduction

How do people make difficult moral trade-offs involving the distribution of benefits, such as when they must decide whether to save twenty people from a moderate harm or instead five other people from a large harm? And how trustworthy are their judgments? The first question is relevant for the development of descriptive theories of social choice and political science, since people's moral judgments will, at least on some occasions, influence which policies have public support. For example, debates about the use of public resources for health are often informed by surveys of the public's views on such trade-offs, and priority-setting policies which lack support will often be withdrawn (see, e.g., Ubel et al. 1996). The second question is relevant for normative theorizing about distributive morality. Following the use of the method of "reflective equilibrium," philosophers commonly test moral principles by their conformity with people's case judgments, or "intuitions" (Rawls 1999: 40-6; Daniels 2013). For this project, it is important to establish under which circumstances people's case judgments are suitable for this purpose because they are likely to reflect a reasonable and considered set of values and when, instead, they are not to be depended upon, because they are likely to be a response to morally irrelevant factors or the result of a biased assessment of morally relevant factors. Naturally, the two questions—about the determinants of our case judgments and their trustworthiness—are related. For one way to argue that particular case judgments should not be trusted is to show that they are merely the upshot of people's use of an undependable mental short-cut or heuristic. Indeed, there is a fast-growing body of research by psychologists and philosophers which casts doubt on moral case judgments in precisely this manner, to the extent that some speak of a "wave of skepticism emanating from the social sciences about the role of intuitive judgments in ethical theory" (Dreisbach and Guevara 2017: 1). This scepticism has been challenged,

however, by other psychologists and philosophers, who have offered what they see as vindicatory explanations of the judgments in question as reflective of attractive moral principles and values.[1]

In this paper, we aim to contribute to these empirical and normative debates regarding people's judgments about distributive justice. We focus on cases in which individuals take on the role of a decision-maker about the use of public health care resources and are asked to balance the number of people they can save against the magnitude of the harm from which they can save them. Subjects presumably hold values that could inform a theory of justice for such cases, which is why social scientists conduct surveys and philosophers devise thought experiments to uncover them. However, subjects are unlikely to have at the ready a fully developed theory of justice to decide them. Moreover, they are unlikely to have much experience with trading off the magnitude of harm prevented against the number of people saved from harm. They can therefore be expected to find some such trade-offs difficult to make.

Some descriptive theories of choice hold that when faced with challenging choices between two-dimensional alternatives, a substantial share of people first see if they can use a heuristic consisting of a simple rule (or a set of such rules, sequentially applied) to make a choice without explicitly trading off dimensions against each other (Tversky 1972; Brandstätter et al. 2006; Manzini and Mariotti 2007; Drechsler et al. 2014; Tserenjigmid 2015). In this paper, we investigate the use of one such heuristic, known as "similarity-based

[1] For references and critical discussion, a good place to start is the journal *Ethics'* symposium issue on experiment and intuition in ethics (Vol. 124). See also Singer (2005); Sunstein (2005) and the associated open peer commentary; and Sinnott-Armstrong (2008).

decision-making" (Tversky 1969, Rubinstein 1988). Subjects who use this heuristic decide as follows when faced with a pair of two-dimensional alternatives. If the alternatives are similar along one dimension and dissimilar along another, they choose the alternative that is better along the dissimilar dimension.

We report an experiment designed to test for use of this heuristic. Our results suggest that, in our experiment, somewhat in excess of 40% of subjects employ it. Moreover, use of this heuristic induces individual and collective choices that are inconsistent with both formal theories of rationality and all standard, substantive theories of distributive justice. We argue that these results indicate the untrustworthiness of moral judgments in cases that elicit similarity-based decision-making.

We proceed as follows. In section 1, we describe the similarity heuristic in more detail, review evidence of its use and indicate our intended contributions. In section 2, we describe the idea underlying our experiment. In section 3, we describe our methods. In section 4, we discuss our results. In section 5, we consider the implications our findings for other studies of people's attitudes and for philosophical thought experiments. In section 6, we conclude.

1. Similarity-based decision-making

A general hypothesis about the role of similarity in a pairwise choice between multi-dimensional alternatives runs as follows. Dimensions along which alternatives appear similar will receive less attention and so receive less weight, while dimensions along which they appear dissimilar will capture attention and so receive greater weight (Mellers and Biagini 1994; Goldstone et al. 1997; Dhar et al. 1999; Köszegi and Szeidl 2013). Here, we focus on the following version of this general hypothesis (Tversky 1969; Rubinstein 1988).

Stage 1: The decision-maker looks for dominance. If the first alternative is at least as good as the second along both dimensions and better on at least one, then the first alternative is chosen.

Stage 2: If Stage 1 does not yield a verdict, the decision-maker compares each dimension separately, looking for similarities. If they perceive similarity in one dimension only, they prefer the alternative that is superior along the dissimilar dimension.

Stage 3: If neither Stage 1 nor Stage 2 yields a verdict, the choice is made using an unspecified different criterion.

There are several reasons that one might use this heuristic. First, because it checks for dominance, it avoids errors that might occur in an overall evaluation of each alternative in isolation. If such overall evaluation were imprecise, it would sometimes select an alternative that was slightly worse on both dimensions, in violation of dominance.

Second, the heuristic draws on readily available and easily evaluable information. Similarity appears to be among the features of objects and alternatives that are routinely and automatically registered by the perceptual system (Engel and Wang 2011). People also appear to find it easier to evaluate differences than absolute magnitudes (Tversky and Kahneman 1979; 1983).

Third, the procedure capitalizes on the fact that intra-dimensional evaluation is relatively simple, because it involves comparisons between features of alternatives that are expressed in the same units (Tversky 1969). Subjects may also lack settled judgments about how to balance a loss on one dimension against a gain in another, which gives them reason

to use a heuristic that side-steps such trade-offs (Tversky 1972; Brandstätter et al. 2006; Manzini and Mariotti 2007; Tserenjigmid 2015).

Notwithstanding these advantages, use of the similarity heuristic may yield choices that are a mere artefact of the choice procedure. It may also lead to violations of principles of rational choice. For example, it can lead to violations of transitivity of strict preference—the requirement that if a decision-maker has a strict preference for alternative *A* over *B,* and for *B* over *C*, then they must strictly prefer *A* to *C*. Suppose that *A* is worse than and similar to *B* along the first dimension, and better than and dissimilar to *B* along the second dimension. The similarity heuristic then leads to a preference for *A* over *B.* Further suppose that *B* is worse than and similar to *C* along the first dimension, and better than and dissimilar to *C* along the second dimension. The similarity heuristic then leads to a preference for *B* over *C.* Finally, suppose that *A* and *C* are dissimilar along both dimensions and that the first dimension is an important determinant of choice when alternatives differ substantially along it. Then, consistently with use of the similarity heuristic, the subject may prefer *C* to *A* in a pairwise comparison.

Following Tversky (1969), the role of similarity in choice has been studied in many experiments. The vast majority of these focus on elf-interested choices between gambles (Lindman and Lyons 1978; Budescu and Weiss 1987; Mellers et al. 1992; Leland 1994; Raynard 1995; Buschena and Zilberman 1995; 1999; Goldstone et al. 1997; Day and Loomes 2010; Loomes 2010; Regenwetter et al. 2011; Brandstätter and Gussmack 2013; Loomes and Pogrebna 2014), but some also examine the influence of similarity on other choices, including the self-interested trade-off between commuting time and wage (Mellers and Biagini 1994) and self-regarding inter-temporal trade-offs (Rubinstein 2003). The predominant findings are: (i) in a substantial share (e.g., in Tversky's experiments, around

one-third) of subjects, similar dimensions receive less weight in decision-making than dissimilar dimensions; and (ii) these subjects are prone to violating transitivity. (Because subjects are known not to choose deterministically, experiments standardly focus on Weak Stochastic Transitivity, which allows for a random error in the process of choice. In repeated pairwise choices between alternatives, this requires that if the probability of choosing *A* over *B* is greater than half and the probability of choosing *B* over *C* is greater than half, then the probability of choosing *A* over *C* must exceed one-half. In what follows, we also employ this conception of transitivity.)

Our intended contributions lie at the intersection of psychology and theories of distributive justice. As indicated, the vast majority of experiments on similarity-based decision-making involve self-regarding choices.[2] One may reasonably expect people to employ the same heuristics that they use in non-moral decision-making to make moral decisions. However, this expectation is worth testing, for two reasons. First, in the light of the current replication crisis in psychology, an attempt to replicate results regarding the use of this heuristic in a new context has epistemic value (Open Science Collaboration 2015; Diener and Biswas-Diener 2017). Second, as mentioned in our Introduction, there is a lively controversy about attempts to explain moral case judgments as the result of the application of heuristics whose use has been established primarily in non-moral domains. A prominent

[2] To our knowledge, the sole exceptions are one of Tversky's (1969) experiments, which involved pairwise choices between potential university applicants and Mellers' (1982) test of the influence of similarity on fairness judgments.

example is the appeal to so-called "loss aversion"[3] to explain the common idea that, other things equal, harming people is morally worse than failing to benefit them (see, e.g., Kahneman 1994; 2011: 369-70; Baron 1998; Horowitz 1998; Sunstein 2005). This explanation is presented as undermining the probative value of this common judgment, since its proponents regard the baseline against which people judge potential losses a mere framing effect, devoid of moral significance. However, this analysis has been rejected by leading philosophers, who have offered competing, vindicatory explanations of the judgments in question as responsive to morally significant factors (see, e.g., Kamm 2007: 422-49; Nebel 2015; Dreisbach and Guevara 2017). Given such disputes, it is worth establishing whether the similarity heuristic may determine people's case judgments in the domain of distributive justice and whether this gives us reason to doubt these judgments.

2. General idea of the experiment

To answer these questions, we aimed to see if the similarity heuristic could yield choices that are contrary to all leading principles of distributive justice, thereby apparently ruling out an alternative, vindicatory explanation of these judgments. (Our experiment therefore introduces a new type of test of the use of the similarity heuristic. Extant experiments infer the prevalence of similarity-based decision-making from violations of formal principles of rational choice, such as transitivity. While we do so too, we also use violations of substantive principles of distributive justice to diagnose its use.)

[3] Loss aversion is the idea that decisions are coded in relation to a baseline from which losses repel more than gains attract.

Standard theories of distributive justice that respect the Pareto principle range from utilitarianism (which requires maximizing the sum-total of well-being, or well-being, generated) to leximin (which requires maximizing the situation of the least-well-off). All such views except utilitarianism are willing to sacrifice some total well-being for the sake of improving the lot of the worst off. And all such views except leximin are willing to accept some worsening in the situation of the worst off for the sake of a sufficiently large improvement in others' well-being. This is true, for example, of forms of pluralist egalitarianism that care about both reducing inequality and improving total well-being (see Tungodden 2003). It is also true of the view known as prioritarianism, which does not care about inequality itself, but which gives some, non-infinite, extra weight to gains in well-being that take place from a lower level (Adler 2012).

How *do* subjects make these trade-offs? In line with the aversion to making trade-offs mentioned in the introduction, some simply avoid them. For example, in one study, Rodriguez-Miguez and Pinto-Prades (2002) report that 26% of subjects choose on the basis of one characteristic only: either they always save the greater number, even when saving the greater number does not maximize total well-being, or they always save those facing the greatest harm, even when saving the better off would do far more good in aggregate. However, the predominant finding across many studies is that when subjects do make trade-offs, they tend to give substantial, though finite, extra weight to gains in well-being to the less well off (Nord and Johansen 2014). Their moral preferences therefore generally align with the aforementioned pluralist egalitarian or prioritarian theories.

To test the hypothesis that a substantial share of subjects would use the similarity heuristic, we proceeded as follows. We constructed pairs of alternatives that involved a trade-off between the magnitude of harm that people were saved from and the number of

people saved so that: (i) these alternatives would appear similar along the "magnitude of harm averted" dimension but dissimilar along the "number of people saved" dimension; and (ii) choosing to save the more numerous group from the somewhat smaller harm would involve *helping the better off at a cost in total well-being*. On these alternatives, we hypothesized, subjects who used the similarity heuristic would favour the more numerous better off, contrary to all aforementioned standard theories of distributive justice and to the moral preferences generally evinced in surveys.[4]

To establish subjects' preferences when similarity could not determine choice, we also designed "wholly dissimilar" alternatives. Confronted with these alternatives, we conjectured, subjects would tend to help the less numerous worse off, both when this maximized total well-being and when helping the worse off would come at some (modest) cost in total well-being.

The conjectured switch between aiding the more numerous better off in choices between partly similar alternatives and aiding the less numerous worse off otherwise would be explicable neither in terms of standard theories of distributive justice nor by the use of

[4] We chose the "magnitude of harm averted" as the dimension along which some alternatives would appear similar because only by doing so could we construct choices in which the similarity heuristic would select the side that is disfavoured by all standard theories of distributive justice. For consider a choice between saving a greater number of people from a smaller harm or a smaller number of people from a greater harm in which the numbers of people in these two groups appear similar but the harms are dissimilar. In that case, similarity-based decision-making would favour saving the marginally smaller group from the much greater harm. But so would, most likely, all standard theories of justice.

the alternative heuristics that avoid trade-offs mentioned by Rodriguez-Miguez and Pinto-Prades (2002), viz., “always help the greater number” or “always help the worst off.” It would, however, be explained by use of the similarity heuristic.

3. Methods

We recruited 82 subjects (72% students, 28% non-students; 51% male, 49% female) from the subject pool of the Centre for Economic Learning and Social Evolution (ELSE) at University College London. (A test run was done with 15 subjects, but since it showed no problems and no revisions were subsequently made to the experiment, these participants were all included in our sample.)

Participants were sat in separate cubicles at individual computer screens. They were informed that they would be paid a flat fee of £13 (roughly USD 20 at the time) for participating in a 40-minute experiment on making choices in the use of health care resources and that a further £5 (USD 8) would be donated to a health care charity of their choice at the end of the experiment. (A full description of the introduction and questionnaire is available in Appendix 1.)

Participants were informed that they would face a series of choices between two interventions and asked which of the two the National Health Service should prioritise. They were asked to suppose that the people affected were in their mid-thirties and in perfect health until recently, but that they now faced a health problem which diminished their well-being to the indicated level. If left untreated, these people would live the rest of a normal human lifespan with the indicated level of well-being; if treated, they would be returned to perfect health for the remainder of this lifespan. The measure of well-being used was the Health Utilities Index Mark III (Feeny et al. 1995, HUI Inc. 2008). This assigns 0 to death, 1 to

perfect health, and a value in between to life in a state of impaired functioning that is better than death.[5] Subjects were told that it was developed by experts and were given a four-screen tutorial on its meaning. This included a picture of the scale, along with the representative valuation of eight conditions. It was explained that the values assigned to life in these conditions were determined by representative answers on surveys and that these values indicated the typical impact on well-being of a condition, with lower numbers indicating lower well-being. Subjects were also informed that, on this scale, an increment of a given size always did a person just as much good, no matter from what level this increment took place.[6]

After this introduction, subjects were presented with four practice choices. The main experiment consisted of three "rounds" of going through sixteen choices in individually randomized order, for a total of forty-eight choices. (Every choice in the main experiment was therefore made three times, with, on average, fifteen other choices between repetitions.) After they had completed their choices, subjects were asked to offer a written explanation of five of their choices.

---

[5] This index relies on the so-called "standard gamble" (Dolan 2001). If subjects respect the von Neumann-Morgenstern axioms, then it is a measure of von Neumann-Morgenstern well-being.

[6] Such stipulations notwithstanding, subjects may treat well-being scales as if they have diminishing marginal prudential value (Greene and Baron 2001). If this were true of our subjects, this would make it even more difficult to achieve the hypothesized preference for aiding the better off at a cost in total well-being, and so a preference of this kind would be even more strongly indicative of the use of similarity-based decision-making.

Alternatives were displayed as in Figure 1. (The placement of alternatives on the right-hand and left-hand side of the screen was randomized.) In the figure, the solid vertical line and number to the left alongside the top of this line represent the group's health status if untreated (0.95 for alternative *A* and 0.91 for *B*). The dotted line represents the health that would be restored by treatment. A box attached to the top of the solid line contained the number of people in that condition that one can treat (48 for *A* and 27 for *B*). Subjects were told that they could treat only one of the two groups. They did so by clicking on the box with the number of patients in that group and moving it all the way up to full health.

4. Results and discussion

As a test of basic comprehension and attentiveness, we included choice in which one alternative dominated another (*N* versus *O* in Table 1). Seventy-six participants (92.7%) selected the dominant alternative three times out of three. A further three (3.7%) chose the dominant alternative two times out of three. Another three participants (3.7%) chose the dominated alternative two or more times out of three. We exclude these last three subjects from the following analysis. (This exclusion makes no substantial difference to our conclusions.)

4.1 Alternatives that are similar in terms of health gain

Consider the choice between *A* and *B* in Figure 1. We conjectured that in this choice, the alternatives would appear similar along the "health-related well-being gained" dimension, but dissimilar along the "number of people helped" dimension. Use of the similarity heuristic would therefore yield a preference for *A*, despite the fact that *B* would both aid the worst off and yield a small improvement in total well-being. As listed in Table 1, we

constructed a further three alternatives, *C*, *D*, and *E*, each of which was designed to appear similar to its immediate predecessor along the health-related well-being gain dimension and each of which yields somewhat greater total well-being than its predecessor.

As Table 2 reveals, helping the better off at a cost in total well-being is indeed common in the pairwise choices between these alternatives. For *A* versus *B*, *B* versus *C*, and *C* versus *D*, more than half of all subjects favour the better off at least 2 out of the 3 times they were presented with the choice. This implies that, if decisions in these pairwise comparisons were taken by majority voting, our group of subjects would choose contrary to every leading theory of distributive justice.

The exception is *D* versus *E*, in which, at 38%, the preference for aiding the better off is less common than in the other choices between adjacent alternatives in the *A* to *E* sequence. (Statistical tests reported in Appendix 2, Table A2.2 confirm that *D* versus *E* stands apart.) Our explanation for this is that, as we move through this sequence, the absolute difference in (and the ratio of) the number of people saved shrinks to the point that some subjects would perceive *D* and *E* as similar along *both* dimensions. For such subjects, the similarity heuristic does not mandate a choice at Stage 2. Instead, it moves to Stage 3, at which we conjectured that subjects would display preferences in line with standard theories of distributive justice, all of which mandate aiding the worse off in this choice.

Because the gap in health gain between non-adjacent alternatives in the set *A* to *E* is larger than between adjacent alternatives, the former are more likely to look dissimilar along both dimensions. A key prediction of our hypothesis is therefore that subjects would be less likely to aid the better off (at a cost in total well-being) in pairwise choices between non-adjacent alternatives than in choices between adjacent alternatives. As Figure 2 shows,

this is indeed what occurred. To establish whether this difference in the rate of aiding the better off is statistically significant, we consider the comparisons listed in Table 3. The underlined numbers in the top-right corner of each comparison indicate the share of subjects who engage in the predicted switching from aiding the better off in a choice between adjacent alternatives to aiding the worse off in a choice between non-adjacent alternatives. (For example, 22.8% of subjects both aid the better off in *A* versus *B* and aid the worse off in *A* versus *C.*) There is no comparable shift in the opposite direction. (For example, only 6.3% of subjects aid the worse off in *A* versus *B* and the better off in *A* versus *C*.) We use McNemar's exact test to calculate the probability of these results given the null hypothesis that the answers to two different pairwise choices are random draws from the same binomial distribution. (For discussion of this test, see Appendix 2.) The grey box in the middle of each comparison in Table 3 reports the results. The differences between adjacent and non-adjacent choices are significant throughout.

This switch from favouring the better off in a choice between adjacent alternatives to favouring the worse off in a choice between non-adjacent alternatives should make it more likely that subjects will display a particular type of intransitive preferences. For example, subjects using the similarity heuristic would favour *A* over *B* and *B* over *C*. But if they also found *A* wholly dissimilar to *C*, they would, in line with standard theories of distributive justice, prefer *C* to *A*, violating transitivity. Of course, such intransitivities could also arise through a subject simply having a fixed probability of making an error (not choosing in line with their judgments) in any given choice. But intransitivities due to such "trembling hand" errors would be equally likely to be of the kind explicable by the similarity heuristic (e.g., choosing *A* over *B* over *C* over *A*) as of a kind not so explicable (e.g. choosing

C over *B* over *A* over *C*). If a large share of subjects employ the similarity heuristic, then we should observe intransitivities of the former kind more often.

To test whether the prevalence of intransitivities is best explained in terms of random error or instead in terms of use of the similarity heuristic, we therefore divide our subjects into three groups: those who do not make intransitive choices (59.5%), those who make intransitive choices in a manner explicable in terms of use of the similarity heuristic (35.4%), and those who make intransitive choices that are not so explicable (5.1%). As Table 4 shows, we can confidently reject the hypothesis that the latter two groups are equally large.

Intransitivities of the kind induced by similarity-based decision-making are also manifest at the group level. As Table 2 reveals, pairwise majority voting yields a group preference for *A* over *B*, *B* over *C*, and *C* over *D*, but it also a preference for *C* over *A*, *D* over *A*, and *B* over *D*, yielding three intransitive cycles.

4.2 Wholly dissimilar alternatives

Despite the larger gap in terms of health gain, some subjects might still have found some non-adjacent alternatives in the set *A* through *E* similar along the health-related well-being gain dimension. After all, the difference in well-being gain between, say, *A* and *C* is only 0.08. We therefore constructed further choices between wholly dissimilar alternatives, all listed in Table 1. Each of the pairwise choices *R* versus *S*, *T* versus *U*, and *V* versus *W* involves a stark choice between helping a substantially smaller, substantially worse off group and helping a much larger, much better off group. In each of these case, helping the worse off yields somewhat greater well-being. Figure 3 reveals that, as predicted, aiding the better off at a cost in total well-being is indeed much more frequent in choices among partly similar

alternatives than in choices among wholly dissimilar alternatives. (Our analysis in Appendix 2, Table A2.3 confirms that this difference is statistically significant.)

To test our conjecture that a substantial number of subjects will both aid the better off at cost in total well-being in choices between partly similar alternatives and aid the worst off at a cost in total well-being in choices between wholly dissimilar alternatives, we constructed *G* versus *F* and *Q* versus *P*. Table 5 shows that our evidence supports this conjecture. For example, a striking 49.4% of all subjects shift from aiding the better off in the choice between *A* and *B* to aiding the worse off in the choice between *G* and *F.* (Only 5.1% switch in the opposite direction.) These differences are statistically significant at the 1% level for all pairs but one.

4.3 Subjects' decision rules

We shall now examine individual-level data. We start by matching individuals with the decision rule that best represents their choices. In doing so, we note that the similarity heuristic is consistent with a wider range of choices than the other decision rules under examination, because it allows for individual-level variability in perceptions of similarity. This flexibility may give the similarity heuristic an "unfair advantage" over other decision rules. We attenuate this problem as follows. We first report an analysis which allows limited variability in individuals' perceptions of similarity. We then consider how robust our findings are by imposing the same perceptions of similarity on all subjects.

Our first test permits only the following two types of perceptions of similarity:

Adjacent Only: All adjacent alternatives in the *A* through *E* sequence, and only these alternatives, are similar along the well-being gain dimension;

Two Steps Only: All alternatives that are no more than one step apart in the *A* through *E* sequence, and only these alternatives, are similar along the well-being gain dimension.

Moreover, we do not consider data from *D* versus *E*, on which the similarity heuristic would have an unfair advantage because it is consistent with either choice, since *D* and *E* may be regarded as similar along the well-being gain dimension only (in which case it predicts that *D* is chosen), or along both dimensions (in which case it predicts that *E* is chosen). We also do not consider *N* versus *O*, since this was a mere basic comprehension test. Since each subject confronted each of the remaining fourteen comparisons three times, this yields 42 data points for each individual. We then assign each individual to the decision rule that gets the largest share of these choices right. The second column of Table 6 displays the results. It indicates that the similarity heuristic is the most common decision rule, with some form of it being the best description of 41.8% of the sample. It also reveals that almost all of those who use the similarity heuristic are prepared to sacrifice total well-being for the sake of the worse off when alternatives are wholly dissimilar. The second-most common decision rule (the uniquely best match for 35.4% and tied for best for a further 5.1%) is to always help the worst off. A small minority (12.7%) is best described as always saving the greater number; an even smaller minority (5.1%, or 10.1% if one counts ties) is best described as maximizing total well-being.

The third column of Table 6 indicates how well these decision rules fit the choices of the subjects matched with them. The similarity heuristic fits its matched subject population reasonably well, with a "success rate" of 78.5%. This heuristic adds substantially to our ability to predict the choices of the subjects matched with it—on average, if we could not

use this heuristic to describe their choices, our success rate at describing them would drop by 11.6%.

As a robustness check, we also considered a version of the similarity heuristic that imposes the uniform perception of (dis)similarity inherent in the aforementioned Adjacent Only rule on all subjects. This is a demanding test of this heuristic, since some diversity in individual perceptions of similarity is to be expected. As detailed in Appendix 2, Table A2.8, this imposition somewhat lowers the share of subjects whose behaviour best matches the similarity heuristic to 36.7%. Moreover, all but one of these subjects give priority to the worst off in choices between wholly dissimilar alternatives. This means that even if we remove all "flexibility" from the similarity heuristic, it is the best fit for 35.4% of subjects, placing it roughly on a par with "always aid the worse off." In sum, the results of our robustness test support the idea that the similarity heuristic is used by around two-fifths of subjects.

Further evidence can be gleaned from subjects' written explanations of five of their choices, which they completed at the end of the experiment. We pigeonholed each of their answers using one of the six categories listed in the first column of Table 7. To illustrate this categorization, consider the following examples of subjects' explanations of their choices in *A* versus *B*. (The categorization was checked by a coder unfamiliar with the study's hypotheses;[7] subjects' complete answers and our categorization can be accessed in Appendix 3.)

S44 offered the following explanation of their preference for *A*:

[7] The independent coder, Carl Runge, was a colleague with experience in coding survey research. They coded all subjects' description of their responses in ignorance of the coding

"Because it helped 21 more people & there was only 0.04 difference in severity of problem."

We categorized this answer as displaying evidence of use of the similarity heuristic.

S47 explained their preference for *B* as follows:

"48 people are almost fine. 27 are worse off; so they should be helped."

We categorized this answer as expressing a special concern for the worse off.

S60 offered the following explanation for their preference for *A*:

"The more number of people to treat [sic]."

We categorized this answer as indicating adherence to a rule requiring saving the greater number.

S41 explained their preference for *B* over *A* as follows:

"the total gain is more because 27*0.09>0.05*48."

This answer was categorized as expressing adherence to the rule of maximizing total well-being.

S81, whose choices expressed a preference for *A*, wrote:

"other people in reasonable health."

This answer was categorized as not rationalizing the choice in question, since it is too terse to serve as a justification. (Other reasons for placing responses in this category were offering reasons that justified choices that differed from the subjects' actual choices, or not answering the question.)

---

done by one of the authors. The two codings were compared and any disagreements were resolved through discussion. Both coders were fully satisfied with the result.

Moreover, it turned out that, especially in *C* versus *D*, a small number of subjects appealed *not* to similarity in terms of well-being gain, but rather similarity in terms of number of people treated. They then decided on the basis of the well-being gain dimension. For example, S45, whose choices expressed a preference for *D*, wrote:

> "more health gain in second state [*D*] and not big difference in number of people treated."

Several results are worth highlighting. First, for choices between alternatives that we hypothesized would be perceived as similar along the well-being gain dimension only, the most frequently offered explanation involves this similarity. Second, in choices that we hypothesized would be seen as wholly dissimilar, concern for the worse off predominates as an explanation. Third, this increase in attention to the worse off is almost entirely due to subjects who switch from appealing to similarity along the gain dimension to justify their choice to appealing to the fate of the worse off. Finally, other rationales, including the aim of maximizing total well-being, are infrequently invoked.

Our individual-level data therefore helps assess a hypothesis raised by a number of commentators, which is that a substantial share of subjects aim to maximize total well-being throughout, but simply make errors in estimating the alternative with the highest total well-being in pairwise choices between adjacent alternatives in the *A* through *E* sequence. These errors, so this hypothesis goes, are committed because the total is difficult to calculate and the difference in total well-being between the alternatives is small.[8]

[8] This hypothesis was raised by Joseph Millum, Antonio Cabrales, and seminar audiences. It gains indirect support from the finding in Arieli et al. (2011) that subjects were more likely

We note that this "people are error-prone utilitarians" hypothesis is compatible with the idea that subjects use the similarity heuristic when they have difficulty engaging in the "holistic" evaluation of alternatives. For it is consistent with the idea that subjects use this heuristic to estimate which alternative maximizes total well-being when calculating this total is demanding. Nonetheless, if correct, it would conflict with our idea, mentioned in the introduction, that in cases of the kind under consideration, many subjects do not yet have a fully articulated theory of distributive justice which they are trying to apply. Our findings, however, offer very little support for the "people are error-prone utilitarians" hypothesis. As we have seen, utilitarianism best fits only a handful of people's choices (see Table 6). This is because in *G* versus *F* and *Q* versus *P*, the vast majority chose to aid the worse off at a cost in total well-being. (It is noteworthy that these are choices in which total well-being was relatively easy to calculate.) Finally, as the final column in Table 7 reveals, in only 2.5% of all cases in which a subject invoked similarity as a rationale for aiding the better off did they also invoke a utilitarian rationale for their choices between wholly dissimilar alternatives.

In sum, subjects' accounts of their choices confirm the conclusions we drew from our analysis of individual-level choice data. Both types of evidence indicate that, with roughly two-fifths employing it, the similarity heuristic is the most commonly used decision rule (closely followed by special concern for the worse off). Moreover, both subjects' choices and their proffered rationales indicate that the vast majority of individuals who employ the similarity heuristic choose on the basis of concern for the worse off when faced with wholly dissimilar alternatives.

---

to engage in separate evaluation of the probability and prize dimensions of gambles when the expected monetary value of the gamble was difficult to compute.

4.4 Limitations

We shall now comment on some limitations of our study. As noted (at the start of Section 3), nearly three-quarters of our subjects were students at a highly selective university; this limits the degree to which one can generalize from our findings to the population at large. Nonetheless, the fact that a simple decision rule that can generate irrational choice behaviour is common among people selected for and trained in abstract, analytical thought suggests that the use of this heuristic may also be widespread in populations without such training.

Further limitations relate to our presentation of the alternatives. As can be seen in Figure 1, the numbers for conditions on the health-related well-being index were depicted in a substantially smaller font than the number of people saved. This may raise the concern that the similarity-induced underweighting of the well-being dimension was partly caused by de-emphasizing this dimension. However, several elements of our presentation of the alternatives mitigate this worry. The information about the well-being dimension was presented not only through the numbers given, but also through the visual representation of the vertical line, which occupied a large part of the depiction of each alternative. Moreover, the well-being that could be added was distinguished by a dotted line, which changed to a solid line as the subject "moved up" the box with the number of people saved from harm, drawing attention to this potential increase. These aspects, along with the fact that the meaning of the well-being measure was discussed across four introductory screens, highlighted the well-being dimension of each alternative. Moreover, if our presentation had underemphasized this dimension, then one would have expected a tendency to underweight the well-being dimension in all decisions. To the contrary, our results in Tables

6 and 7 show that the well-being dimension received very substantial decision weight from the vast majority in all choices except those designed to display similarity along the well-being dimension.

In general, we of course acknowledge that particularities of our arrangement of the alternatives will have influenced subjects' perceptions of similarity and therefore their choices. For example, the visual representation of the well-being scale, along with the fact that numbers were given for the starting-point and end-point of health improvements but not for the size of the well-being gain, may have contributed to making adjacent alternatives in the sequence *A* through *E* seem similar along the well-being dimension. Such dependency of people's judgments on framing is par for the course: all tests of the use of this heuristic rely on perceptions of similarity that are induced by a particular depiction of the alternatives. (For example, in Tversky's 1969 experiment involving preferences over gambles, the prizes—which were designed to be perceived as dissimilar from each other—were precise monetary amounts, whereas probabilities of winning—which were designed to be perceived as similar in at least some choices—were presented as coloured areas of a circle without numerical values.) Nonetheless, we submit that our experiment adds credibility to the general hypothesis that a substantial share of people can be expected to use this heuristic in moral decision-making when (a) they face pairwise choices between multi-dimensional alternatives, neither of which is obviously superior to the other and for which trade-offs between dimensions are difficult to make and (b) the alternatives are presented in a way that generates perceptions of similarity along a key dimension or dimensions, but dissimilarity along other dimensions.

5. Implications for surveys and thought experiments

We shall now connect our findings to the broader debate, referenced in the Introduction, on the extent to which psychological research on the causes of people's intuitive moral judgments undermines their status as reflective of people's deeply held, presumptively reasonable values. While we do not side with those such as Singer (2005) who believe that the psychology of intuitive judgment offers grounds for a wholesale dismissal of these case judgments, we do take our findings to justify scepticism regarding the probative value of people's distributive judgments when these are a product of the similarity heuristic. After all, in our experiment, people's judgments when influenced by this heuristic appear to reflect neither the distributive principles they themselves apply in non-similarity cases nor any reasonable theory of distributive justice.

This scepticism, while limited, nonetheless extends to a range of cases in social science and philosophy. By way of illustration, we shall concentrate on people's case judgments in an area of research which, like our study, involves inter-personal trade-offs, but which, unlike our study, concerns cases in which one can either do a great deal of good for a small group of badly off people or instead improve the lot of a large number of equally badly off people to a small degree.[9]

[9] We note, however, that our scepticism also extends to valuations expressed in some intra-personal trade-offs. For example, Kvamme et al. (2010) find that individuals were willing to pay less per week of extra life for extending an imagined life expectancy of ten years by a short period (e.g., one week, so that they would live for another ten years and a week) than for a substantial period (e.g., one year, so that they would live for another eleven years). If paying some significant amount of money is perceived as dissimilar to paying nothing, and

A common finding in this area is that a substantial share of study participants believe that one ought to provide each member of the small group with the great benefit. For example, Choudry et al. (1997, Table 4) finds that a substantial majority of their sample of senior Canadian health professionals preferred providing twenty years of additional life expectancy to a group of 500 cancer patients rather than one year of additional life expectancy to a different group of 10,000 cancer patients. The additional time alive was posited to be of equal quality for all patients, who were also assumed to be equal in other respects (e.g., age at time of treatment). The preference for treating the 500 is therefore not readily explained by participants' adherence to utilitarianism (which, if one assumes every additional year alive is of equal well-being value, would counsel indifference). Moreover, since, among the patients considered, aiding the 500 makes them much better off than the 10,000, who would, if aided instead, be only somewhat better off than the 500,

(as one readily imagines), living for another ten years and a week is perceived of as similar to living for another ten years, but living for another eleven years is not so perceived, then the similarity heuristic predicts their finding. We submit that this undermines its usefulness as a measure of their participants' true valuation of more time alive.

Other cases abound in philosophy. Arguably, the famous case of the "self-torturer" (Quinn 1990) implicitly prompts people to use the similarity heuristic when making intra-personal trade-offs between money received and the intensity of pain undergone (Voorhoeve and Binmore 2006) as do cases proposed in Rachels (1998) and Temkin (2012) involving trade-offs between intensity of pain undergone and its duration (Voorhoeve 2013). If so, then these cases, too, are not sources of dependable intuitions.

aiding the smaller group goes against all theories of justice that give some additional weight to benefits to the worse off.

Likewise, Rodriguez-Miguez and Pinto-Prades (2002) find that the general tendency to give some extra weight to benefits to the worse off disappears when subjects must choose between giving a large benefit to a small number of badly off individuals (e.g., giving each of two young people facing death fifty years' additional life in good health) and a small individual benefit to many badly off individuals (e.g., giving each of one hundred young people facing death one year of additional life in good health). In such cases, they find that participants tend to offer the large benefit to the few, even when spreading the benefits among many individuals would reduce inequality at no cost in total well-being. (This result is replicated in Hukin and Tsuchiya 2005; for discussion, see also Gaertner and Schokkaert 2012: chap. 5.) These findings are in line with the intuitive judgments that Temkin (2005) reports making (and attempts to elicit from readers) in analogous thought experiments.

Choudry et al. (1997) and Temkin (2005) see people's responses to such cases as offering support for the idea that policy-makers have reason to concentrate a given sum of benefits among a few rather than spread these benefits among a great many people. We suggest an alternative understanding of these results. For people may conceive of such cases as a choice between two multi-dimensional alternatives, where each person's well-being level is one dimension (Tserenjigmid 2015). So conceived, the alternatives are dissimilar in terms of the well-being of each person in the smaller group, but similar in terms of the well-being of each person in the larger group. The similarity heuristic then yields the

verdict that one ought to benefit the smaller group.[10] Our findings, together with previous research on the similarity heuristic, suggest that this explanation has some plausibility. It is also worth noting that the similarity hypothesis has an explanatory advantage over the idea that people have a genuine preference to concentrate a given sum of benefits among a few rather than disperse it among many. For the latter hypothesis leaves unexplained our finding that subjects chose to provide *smaller* individual benefits to a larger, better-off group at a cost in total well-being. The similarity heuristic, by contrast, has the potential to provide a unified explanation of the violation of standard principles of distributive justice in our experiment and in Choudry et al. (1997), Rodriguez-Miguez and Pinto-Prades (2002), and Temkin (2005). Moreover, as we have argued, our hypothesis has normative relevance. If the similarity heuristic is indeed behind these judgments, then we have reason to discount rather than respect them.

6. Conclusion

We have examined how people choose when they must either save a larger number of people from a smaller harm, or, instead, save a smaller number of people from a greater

[10] Of course, this is not the only way in which subjects may conceive of these alternatives. Another way is see them as two-dimensional alternatives, with the first dimension being individual harm averted and the second dimension the number of people saved from this harm. This way of evaluating the alternatives would yield the judgment that they are dissimilar along both dimensions, and hence preclude using the similarity heuristic to resolve the choice. Such dependence of the verdict yielded by a heuristic on the way alternatives are perceived is of course common in the literature on framing effects.

harm (where harm is measured by a loss in health-related well-being). We have documented a remarkable shift in subjects' decisions. In choices between alternatives that appear similar only along the "magnitude of harm prevented" dimension, a majority of subjects help the more numerous better off at a cost in total well-being. By contrast, in choices between wholly dissimilar alternatives, a vast majority of subjects help the less numerous worse off, even when this comes at a cost in total well-being. This shift leads to violations of transitivity at the individual and collective level. We have argued that these patterns of choice are best explained by widespread use of the similarity heuristic; indeed, both individual-level choice data and subjects' written accounts of their choices indicate that somewhat in excess of 40% of our subjects employ this heuristic.

In our experiment, these subjects' choices do not express a consistent set of values. Moreover, their similarity-induced choices, taken separately, are at odds with every standard theory of distributive justice and, taken as a set, are often at odds with formal requirements of rational choice. To us, these facts indicate that their similarity-induced choices are mere contrivances of the choice situation. Indeed, it seems that when deciding between alternatives that are similar only in terms of the harm from which individuals can be saved, similarity-based reasoning leads subjects to systematically underweight this harm's importance.

We have also argued that other striking results in empirical social choice and moral philosophy, such as the finding that individuals prefer to concentrate rather than disperse a given sum of benefits even when this runs contrary to all standard theories of justice, may well be explained by subjects' use of the similarity heuristic and that these case judgments should therefore be treated as untrustworthy. More generally, we conclude that cases which prompt use of the similarity heuristic are suitable neither for surveying the public to

uncover its deeply held moral values nor for testing distributive theories. Our recommendation to social scientists and philosophers engaged in these projects is therefore to construct choices that do not induce similarity-based decision-making.

References

Adler, M. 2012. *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis.* Oxford: Oxford University Press.

Arieli, A., Y. Ben-Ami and A. Rubinstein. 2011. Tracking decision-makers under uncertainty. *American Economic Journal: Microeconomics,* 3: 68—76. DOI: 10.1257/mic.3.4.68

Baron, J. 1998. *Judgment Misguided: Intuition and Error in Public Decision Making*. Oxford: Oxford University Press.

Brandstätter, E. and E. Gussmack. 2013. The cognitive processes underlying risky choice. *Journal of Behavioral Decision Making,* 26: 185–197. DOI: 10.1002/bdm.1752

Brandstätter, E., G. Gigerenzer and R. Hertwig. 2006. The priority heuristic: Making choices without trade-offs. *Psychological Review* 113: 409—32. DOI: 10.1037/0033-295X.113.2.409

Budescu, D. and W. Weiss. 1987. Reflection of transitive and intransitive preferences: A test of prospect theory. *Organizational Behaviour and Human Decision Processes* 39: 184—202. DOI: 10.1016/0749-5978(87)90037-9

Buschena, D. and D. Zilberman. 1995. Performance of the similarity hypothesis relative to existing models of risky choice. *Journal of Risk and Uncertainty* 11: 233—62. DOI: 10.1007/BF01207788

Buschena, D. and D. Zilberman. 1999. Testing the effects of similarity on risky choice: Implications for violations of expected utility. *Theory and Decision* 46: 251—76. DOI: 10.1023/A:1005066504527

Choudhry, N., P. Slaughter, K. Sykora and C.D. Naylor. 1997. Distributional dilemmas in health policy: large benefits for a few or smaller benefits for many? *Journal of Health Services Research and Policy* 2(4): 212—16.

Daniels, N. 2013. Reflective equilibrium. *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/reflective-equilibrium/>.

Day, B. and G. Loomes. 2010. Conflicting violations of transitivity and where they may lead us. *Theory and Decision* 68: 233–242. DOI: 10.1007/s11238-009-9139-1

Dhar, R., S.M. Nowlis and S.J. Sherma. 1999. Comparison effects on preference construction. *Journal of Consumer Research* 26: 293—306. DOI: 10.1086/209564

Diener, E. and R. Biswas-Diener. 2017. The replication crisis in psychology. In *Noba Textbook Series: Psychology*, ed. R. Biswas-Diener and E. Diener. Champaign, IL: DEF publishers. DOI:http://noba.to/q4cvydeh

Dolan, P. 2001. Output measures and valuation in health. In *Economic Evaluation in Health Care*, ed. M. Drummond and A. McGuire, 46—67. Oxford: Oxford University Press.

Drechsler, M., K. Katsikopoulos and G. Gigerenzer. 2014. Axiomatizing bounded rationality: the priority heuristic. *Theory and Decision* 77: 183—96. DOI: 10.1007/s11238-013-9393-0

Driesbach, S. and D. Guevara. 2017. The Asian Disease Problem and the ethical implications of Prospect Theory. *Noûs* online early, DOI: 10.1111/nous.12227

Engel, T. and X.-J. Wang. 2011. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *The Journal of Neuroscience* 31(19): 6982–96. DOI:10.1523/JNEUROSCI.6150-10.2011

Feeny, D., W. Furlong, M. Boyle and G.W. Torrance. 1995. Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics* 7: 490—502. DOI: 10.2165/00019053-199507060-00004

Gaertner, W. and E. Schokkaert. 2012. *Empirical Social Choice: Questionnaire-Experimental Studies on Distributive Justice*. Cambridge: Cambridge University Press.

Goldstone, R., D. Medin and J. Halberstadt. 1997. Similarity in context. *Journal of Memory and Cognition* 25: 237—55. DOI: 10.3758/BF03201115

Greene, J. and J. Baron. 2001. Intuitions about declining marginal utility. *Journal of Behavioral Decision Making* 14: 243—55. DOI: 10.1002/bdm.375

Horowitz, T. 1998. Philosophical intuitions and psychological theory. *Ethics* 108: 367–85. DOI: 10.1086/233809

HUI Inc. 2008. *The Health Utilities Index Mark 3*. http://www.healthutilities.com/ [Last accessed October 13, 2016].

Hukin, A. and A. Tsuchiya. 2005. Dispersion vs. concentration of health benefits: preliminary report, unpublished ms. School of Health and Related Research, University of Sheffield.

Kahneman, D. 1994. *The Cognitive Psychology of Consequences and Moral Intuition*. Delivered as a Tanner Lecture on Moral Values, unpublished manuscript.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kamm, F.M. 2007. *Intricate Ethics*. Oxford: Oxford University Press.

Köszegi, B. and A. Szeidl. 2013. A model of focusing in economic choice. *Quarterly Journal of Economics* 128(1): 53—104. DOI: 10.1093/qje/qjs049

Kvamme, M.K., D. Gyrd-Hansen, J.A. Olsena and I.S. Kristiansen. 2010. Increasing marginal utility of small increases in life-expectancy? Results from a population survey. *Journal of Health Economics* 29: 541–548.

Leland, J. 1994. Generalized similarity judgments: An alternative explanation for choice anomalies. *Journal of Risk and Uncertainty* 9: 151—72. DOI: 10.1007/BF01064183

Lindman, H. and J. Lyons. 1978. Stimulus complexity and choice inconsistency among gambles. *Organizational Behavior and Human Performance* 21: 146—59. DOI: 10.1016/0030-5073(78)90046-6

Loomes, G. 2010. Modeling choice and valuation in decision experiments. *Psychological Review* 117(3): 902—24. DOI: 10.1037/a0019807

Loomes, G. and G. Pogrebna. 2014. Testing for independence while allowing for probabilistic choice. *Journal of Risk and Uncertainty* 49: 189–211. DOI: 10.1007/s11166-014-9205-0

Manzini, P. and M. Mariotti. 2007. Sequentially rationalizable choice. *American Economic Review* 97 (5): 1824—39. DOI: 10.1257/aer.97.5.1824

Mellers, B. 1982. Equity judgment: A revision of Aristotelian views. *Journal of Experimental Psychology: General* 111: 242—70. DOI: 10.1037/0096-3445.111.2.242

Mellers, B., S. Chang, M. Birnbaum and L. Ordonez. 1992. Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance* 18: 347—61. DOI: 10.1037/0096-1523.18.2.347

Mellers, B. and K. Biagini. 1994. Similarity and choice. *Psychological Review* 101: 505—18. DOI: 10.1037/0033-295X.101.3.505

Nebel, J. 2015. Status quo bias, rationality, and conservatism about value. *Ethics* 125: 449-76. DOI: 10.1086/678482

Nord, E. and R. Johansen. 2014. Concerns for severity in priority setting in health care: A review of trade-off data in preference studies and implications for societal willingness to pay for a QALY. *Health Policy* 116: 281—8. DOI: 10.1016/j.healthpol.2014.02.009

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251), aac4716, DOI: 10.1126/science.aac4716

Quinn, W. 1990. The puzzle of the self-torturer. *Philosophical Studies* 59: 79—90. DOI: 10.1007/BF00368392

Rachels, S. 1998. Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy* 76: 71—83. DOI: 10.1080/00048409812348201

Ranyard, B. 1995. Reversals of preference between compound and simple risks: The role of editing heuristics. *Journal of Risk and Uncertainty* 11: 159—75.

Rawls, J. 1999. *A Theory of Justice, revised, 2nd edition*. Oxford: Oxford University Press.

Regenwetter, M., J. Dana and C. Davis-Stober. 2011. Transitivity of preferences. *Psychological Review* 118: 42—56. DOI: 10.1037/a0021150

Rodriguez-Miguez, E. and J.-L. Pinto-Prades. 2002. Measuring the social importance of concentration or dispersion of individual health benefits. *Health Economics* 11: 43—53. DOI: 10.1002/hec.643

Rubinstein, A. 1988. Similarity and decision-making under risk (Is there a utility theory resolution to the Allais Paradox?) *Journal of Economic Theory* 46: 145—53. DOI: 10.1016/0022-0531(88)90154-8

Rubinstein, A. 2003. Economics and psychology? The case of hyperbolic discounting. *International Economic Review* 44: 1207—16. DOI: 10.1111/1468-2354.t01-1-00106

Singer, P. 2005. Ethics and intuitions. *The Journal of Ethics* 9: 331-352. DOI 10.1007/s10892-005-3508-y

Sinnott-Armstrong, W., ed., 2008. *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: The MIT Press.

Sunstein, C. 2005. Moral heuristics. *Behavioural and Brain Sciences* 28: 531–73. DOI: 10.1017/S0140525X05000099

Temkin, L. 2005. A new principle of aggregation. *Philosophical Issues* 15: 218-34. DOI: 10.1111/j.1533-6077.2005.00063.x

Temkin, L. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Tserenjigmid, G. 2015. Theory of decisions by intra-dimensional comparisons. *Journal of Economic Theory* 159: 326—38. DOI: 10.1016/j.jet.2015.07.001

Tungodden, B. 2003. The value of equality. *Economics and Philosophy* 19: 1—44. DOI: 10.1017/S0266267103001007

Tversky, A. 1969. Intransitivity of preferences. *Psychological Review* 76: 31—48. DOI: 10.1037/h0026750

Tversky, A. 1972. Elimination by aspects: A theory of choice. *Psychological Review* 79: 281–99. DOI: 10.1037/h0032955

Tversky, A. and D. Kahneman. 1979. Prospect theory. *Econometrica* 47: 263—91. DOI: 10.2307/1914185

Tversky, A. and D. Kahneman. 1983. Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* 90: 293—315. DOI: 10.1037/0033-295X.90.4.293

Ubel, P.A., G. Loewenstein, D. Scanlon and M. Kamlet. 1996. Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. *Medical Decision-Making* 16: 108—16. DOI: 10.1177/0272989X9601600202

Voorhoeve, A. and K. Binmore. 2006. Transitivity, the sorites paradox, and similarity-based decision-making. *Erkenntnis* 64: 101—14. DOI: 10.1007/s10670-005-2373-1

Voorhoeve, A. 2008. Heuristics and biases in a purported counterexample to the acyclicity of 'better than'. *Politics, Philosophy and Economics* 7: 285—99. DOI: 10.1177/1470594X08092104

Voorhoeve, A. 2013. Vaulting intuition: Temkin's critique of transitivity. *Economics and Philosophy* 29: 409—25. DOI: 10.1017/S0266267113000321

Figure 1. A choice between alternatives *A* and *B*.

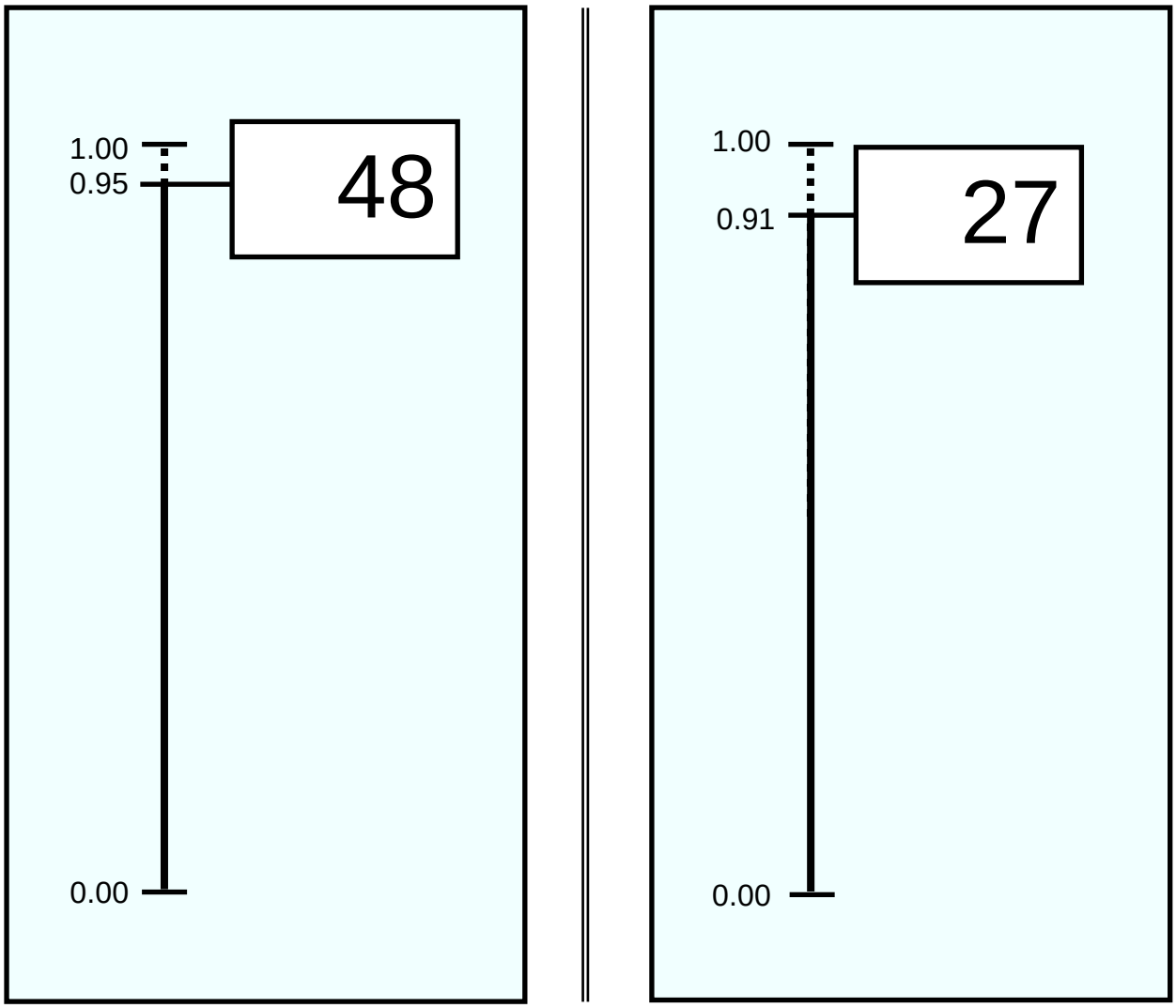


Table 1. All alternatives

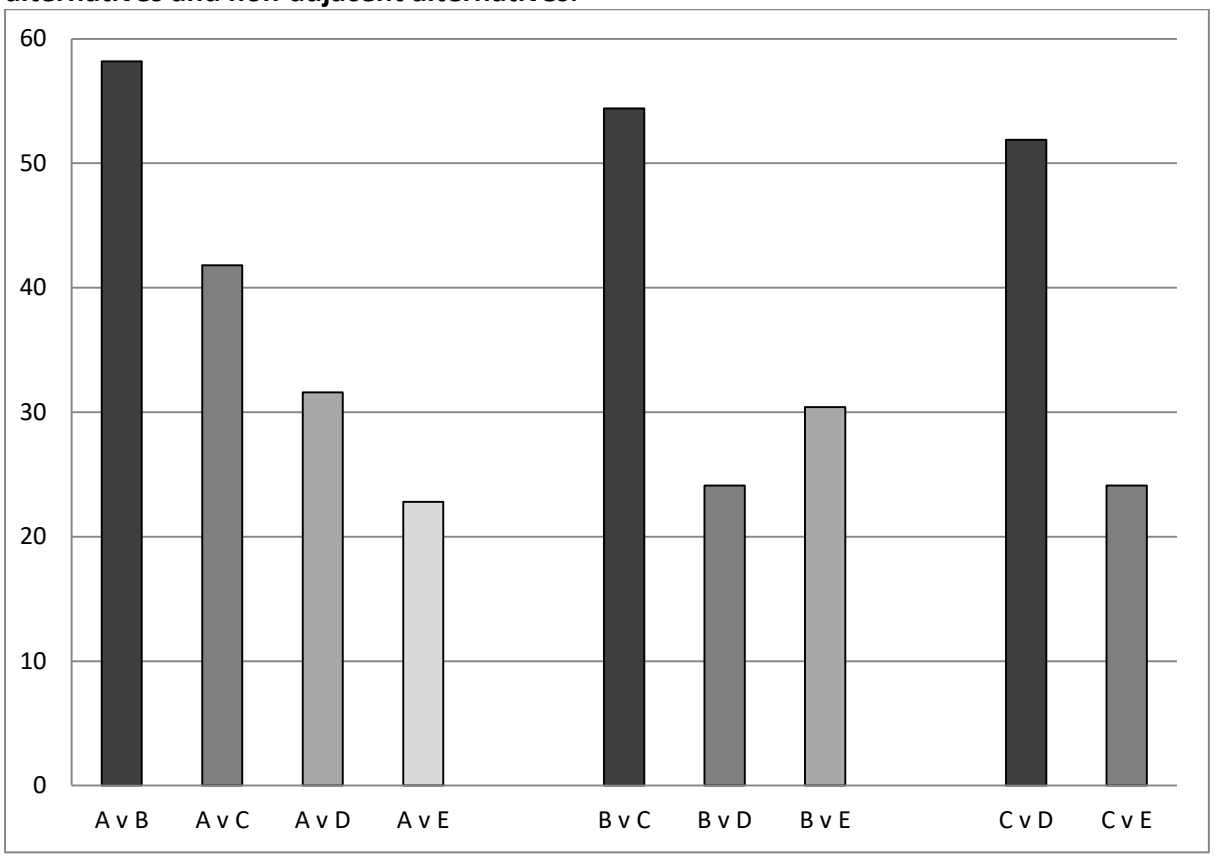
Alternatives		Number of patients	Individual well-being without treatment	Individual gain through treatment	Sum of utilities generated
Similar to adjacent alternative along well-being gain dimension; aiding better off lowers total well-being	*A*	48	0.95	0.05	2.40
	B	27	0.91	0.09	2.43
	C	19	0.87	0.13	2.47
	D	15	0.83	0.17	2.55
	E	12	0.78	0.22	2.64
Note: Subjects choose between all possible pairings from the set *A* through *E*.					
Wholly dissimilar; aiding better off lowers total well-being	*R*	10	0.31	0.69	6.90
	S	20	0.70	0.30	6.00
	T	5	0.26	0.74	3.70
	U	20	0.84	0.16	3.20
	V	5	0.50	0.50	2.50
	W	21	0.90	0.10	2.10
Wholly dissimilar; aiding better off raises total well-being	*F*	10	0.05	0.95	9.50
	G	50	0.80	0.20	10.0
	P	1	0.10	0.90	0.90
	Q	4	0.75	0.25	1.00
Basic comprehension	*N*	31	0.14	0.86	26.70
	O	14	0.35	0.65	9.10

Table 2. Subjects expressing a preference for aiding the better off, in percent

Choices		Percentage of individuals (*n* = 79) expressing preference for better off (≥ 2 times out of 3)
Similar in terms of health gain; aiding better off lowers total well-being	*A* v *B*	58.2
	B v *C*	54.4
	C v *D*	51.9
	D v *E*	38.0
Possible dissimilarity along both dimensions; aiding better off lowers total well-being	*A* v *C* *A* v *D*	41.8 31.6
	B v *D* *B* v *E*	24.1 30.4
	C v *E*	24.1
Wholly dissimilar; aiding better off lowers total well-being	*A* v *E*	22.8
	S v *R*	12.7
	U v *T*	13.9
	W v *V*	13.9
Wholly dissimilar; aiding better off raises total well-being	*G v F*	13.9
	Q v *P*	24.1

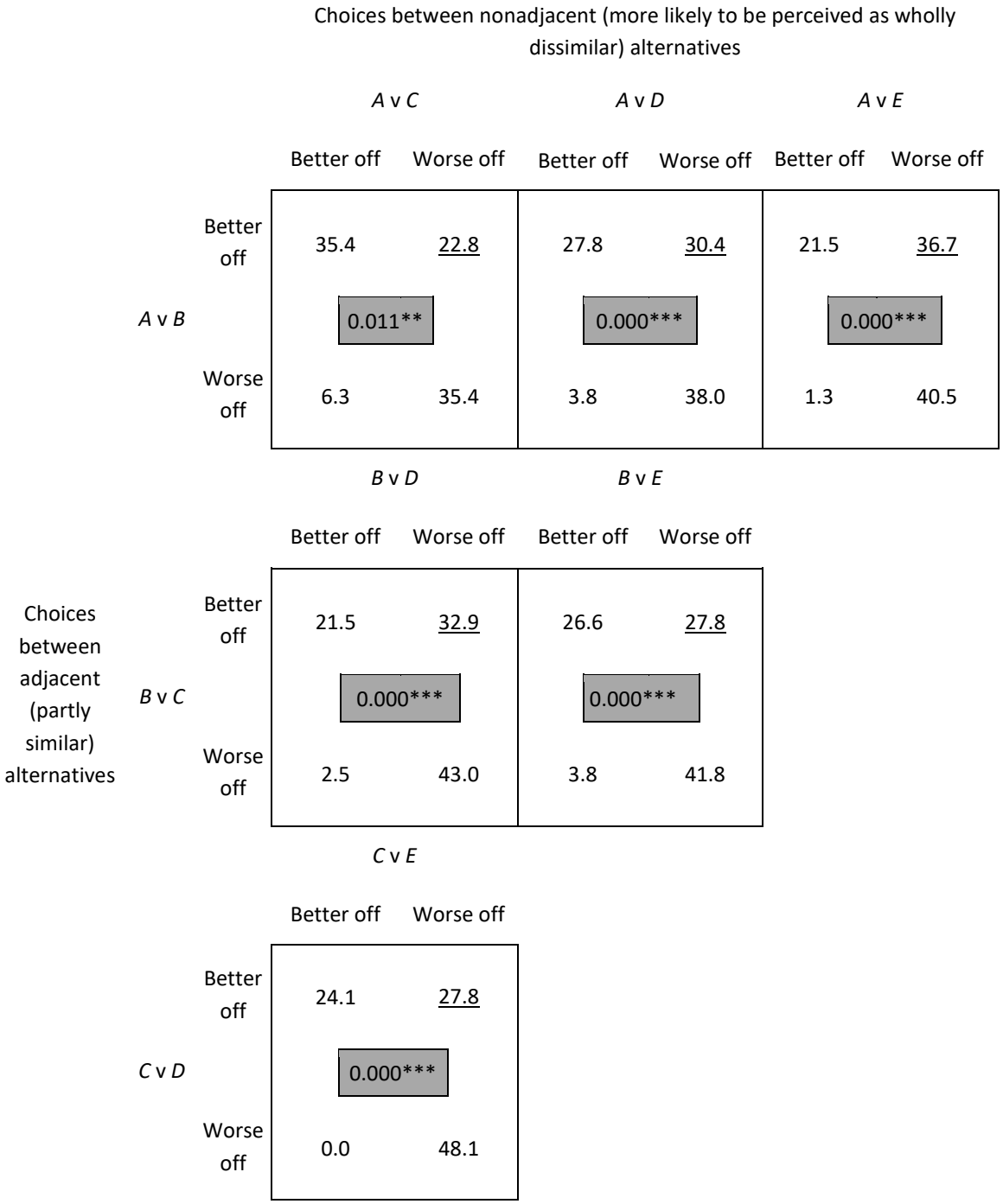
Note: *n* = 79. In the pairwise choices in the second column, the alternative which involves aiding the better off is always listed first. Note also that one can also consider the percentage of all *choices* (rather than subjects) that favour the better off. The resulting pattern is very similar; see Appendix 2, Table A2.1.

Figure 2. Number of subjects (in percent) aiding the better off in choices between adjacent alternatives and non-adjacent alternatives.



Note: *n* = 79. Dark bars indicate choices between adjacent alternatives, which are more likely to be perceived as similar in terms of health gain only. Lighter bars indicate choices between non-adjacent alternatives; these are more likely to be perceived as wholly dissimilar. They are lighter the further apart the alternatives are. Aiding the better off (at a cost in total well-being) is much more frequent among adjacent alternatives than among non-adjacent alternatives.

Table 3. Comparison of choices between adjacent alternatives with choices between nonadjacent alternatives.



Note: *n* = 79. Numbers in the comparisons of distributions across (aiding the better off at a cost in total well-being, aiding the worse off at a gain in total well-being) are percentages of the total sample. Underlined numbers represent the predicted shift from aiding the better off when choosing between adjacent alternatives to aiding the worse off when choosing among non-adjacent alternatives. Numbers in the grey boxes give the probability of obtaining the observed results if the answers come from the same distribution, using McNemar's exact test. We can reject the hypothesis that choices between adjacent alternatives and choices between nonadjacent alternatives come from the same distribution.

** = 5% confidence level.

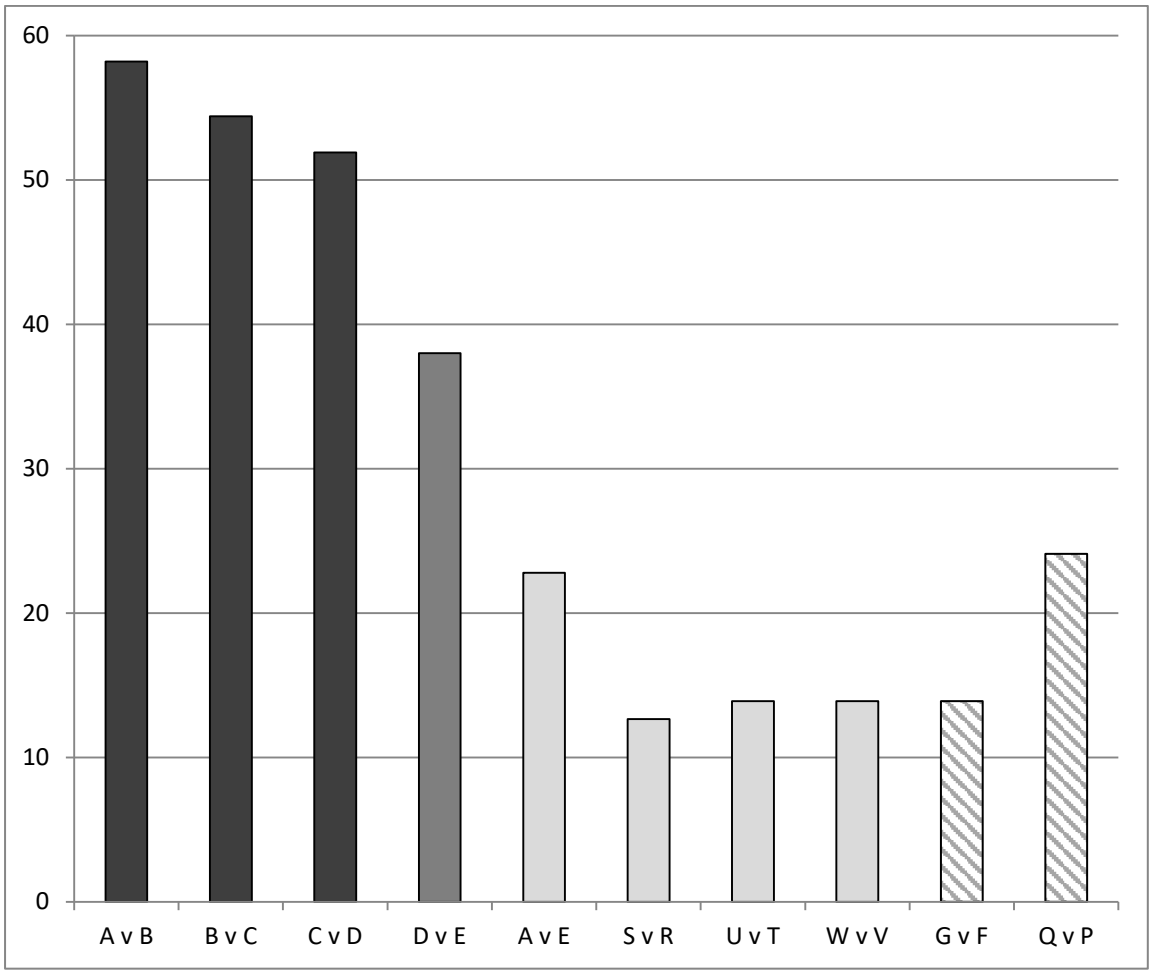
*** = 1% confidence level.

Table 4. Types of intransitivities observed.

	Proportion of subjects, in %	Probability no difference in direction of intransitivity
No intransitivities	59.5	
Intransitivity explicable by similarity	35.4	0.000***
Intransitivity not explicable by similarity	5.1	

Note: n = 79. The final column lists the probability p of obtaining the observed proportions if the probability q of each type of intransitivity were the same. Note that this can be done for any value of q between 0 and 0.5. Therefore, the q that maximises p is chosen. The probability is calculated using equation A2.2 in Appendix 2. In line with our prediction, we can confidently reject the hypothesis that both types of intransitivity are equally likely.
*** = 1% confidence level.

Figure 3. Comparison of the rate of preference for the better off (in percent) in choices between alternatives that are similar in terms of health gain with this rate for choices among wholly dissimilar alternatives.



Note: *n* = 79. Darker bars indicate choices between alternatives that are more likely to involve alternatives that are similar in terms of health gain only; light bars indicate choices between wholly dissimilar alternatives. Bars with an even colouring indicate choices in which aiding the better off decreases total well-being. Patterned bars indicate choices in which aiding the better off increases total well-being.

Two conclusions are apparent. First, giving priority to the better off at a cost in total well-being is much more frequent in choices among alternatives that are similar in terms of health gain. (Appendix 2, Table A2.3 confirms that this difference is statistically significant.) Second, in choices between wholly dissimilar alternatives (*G* v *F* and *Q* v *P*), the vast majority of subjects is prepared to sacrifice total well-being for the sake of the worst off.

Table 5. The shift from aiding the better off to aiding the worse off.

			Wholly dissimilar choices; aiding better off raises total well-being			
			G v F		*Q v P*	
			Better off	Worse off	Better off	Worse off
Similarity in terms of health gain; aiding better off lowers total well-being	*A* v *B*	Better off	8.9	<u>49.4</u>	13.9	<u>44.3</u>
			0.000***		0.000***	
		Worse off	5.1	36.7	10.1	31.6
	B v *C*	Better off	10.1	<u>44.3</u>	13.9	<u>40.5</u>
			0.000***		0.000***	
		Worse off	3.8	41.8	10.1	35.4
	C v *D*	Better off	12.7	<u>39.2</u>	13.9	<u>38.0</u>
			0.000***		0.000***	
		Worse off	1.3	46.8	10.1	38.0
	D v *E*	Better off	7.6	<u>30.4</u>	8.9	<u>29.1</u>
			0.001***		0.090*	
		Worse off	6.3	55.7	15.2	46.8
		Total	**13.9**	**87.1**	**24.1**	**75.9**

Note: *n* = 79. Numbers in the comparisons of distributions across (aiding the better off at a cost in total well-being, aiding the worse off at a cost in total well-being) give percentages of the total sample. Underlined numbers represent the predicted shift. Numbers in the grey boxes give the probability of obtaining the observed results if the answers come from the same distribution according to McNemar's exact test. A large share of subjects switch from aiding the better off at a cost in total well-being when choosing between similar alternatives to aiding the worst off at a cost in total well-being when choosing between wholly dissimilar alternatives.

* = 10% confidence level.

*** = 1% confidence level.

Table 6. Matching subjects with decision rules

Rule	Share (%)	Fit (%)	Fit premium (%)
Similarity heuristic	41.8	78.5	11.6
When no similarity:			
Worst off	40.5	78.9	
Total well-being	1.3	66.7	
Worse off	35.4	88.8	8.3
Greater number	12.7	80.0	23.1
Total well-being	5.1	79.2	4.8
Worse off/total well-being (tie)	5.1	76.8	n.a.

Note: *n*= 79, with 42 choices per person. "Fit" is the share of choices (in those subjects in whose behaviour it fits best) consistent with the rule in question. The "fit premium" is the difference between the share of these subjects' choices explained by the given rule and the share of these subjects' choices explained by the next-best-fitting rule.

Table 7. Subjects' rationales.

Choice / **Rationales**	**Hypothesized to be perceived as similar in terms of health gain**		**Hypothesized to be perceived as wholly dissimilar**			**Subjects who use similarity in health gain switch to the following in wholly dissimilar choices**
	A v *B*	*C* v *D*	*A* v *E*	*G* v *F*	*Q* v *P*	
Similarity in health gain	40.5	41.8	2.5	0.0	0.0	
Similarity in number of people	0.0	7.6	0.0	0.0	2.5	3.0
Worse off	30.4	29.1	63.3	81.0	65.8	66.2
Greater number	10.1	3.8	17.7	11.4	15.2	21.9
Total well-being	5.1	2.5	3.8	1.3	5.1	2.5
No rationale	13.9	15.2	12.7	6.3	11.4	6.5
Total	100.0	100.0	100.0	100.0	100.0	100.0

Note: *n* = 79. Numbers are percentages of all subjects, except the final column, which lists the share of subjects that appealed to similarity to explain at least one of their choices.