The Secret Ingredients to Moral Philosophy: Blood, Sweat, and Tears

On Bad Enough Worst-Case Scenarios in Experimental Approximations of John Rawls' Original Position

Isa Lappalainen

Bachelor Thesis in Political Science

Supervisors: Johan Wejryd & Jonas Hultin Rosenberg

Pages: 37 / Word Count: 13865

Uppsala University

Fall 2018

CONTENTS

INTRODUCTION

"Rawls does not conceive of moral philosophy as depending primarily on the analysis of valid moral argument. Rather, he thinks of a theory of justice as analogous to a theory in empirical science. It has to square with what he calls 'facts', just like, for example, physiological theories. But what are those facts?"

R.M. HARE
(Frohlich & Oppenheimer 1992:20)

In 1971, the publication of *A Theory of Justice* by John Rawls put an end to the 50s and 60s' "dog days of moral and political philosophy" (Freeman 2007:4). Selling over 250000 copies, Rawls shook to the core the two schools of moral philosophy which for two hundred years had dominated the playing field: utilitarianism and intuitionism. With a kantian approach to human agency and reason, and the thought experiment: the original position, Rawls argued for liberal rights and an egalitarian distribution of resources (2007:6).

Rawls was not the first to claim justice to be linked to impartiality (judgments made independently of particular identities or circumstances). The tradition goes back both to kantianism and to Adam Smith who in *The Theory of Moral Sentiments* (1759) made use of an "impartial spectator" (Mongin 2001:150). Because of the impossibility to rid oneself of one's particular identity, however, an impartial observer has, in practice, been thought unattainable. Rawls came up with the original position to circumvent this problem through the creation of certain *conditions* of impartiality. What distributional principles, asked Rawls, would be chosen by self-interested actors whose natural talents, 'conceptions of the good'[1], and positions in society were concealed by a 'veil of ignorance'? He further assumed his agents to be non-altruistic, i.e. not willing to sacrifice their own well-being for that of others, and for the society's economic stage of development to be unknown. Rawls believed the final answer to be equal rights to basic liberties, and the distribution of resources which *most benefits the worst off*. Inequalities, according to Rawls, would only be tolerated if they were to the benefit of the worst off. He called this the maximin principle (Rawls 1999:10-15). The curious mechanism of Rawls' thought experiment is that you, through the insertion of 'self-interestedness', surprisingly, can get 'justice'. Since actors in the original position cannot exploit the situation to cater to any *particular* circumstance, he considered the principles fruits of impartial reasoning and, therefore, principles of justice (Rawls 1999:17).

Unsurprisingly, Rawls' deductions have been met with their fair share of skepticism. And while Rawls was careful to describe the original position as a "purely hypothetical situation"

[1] Life philosophies

(1999:104) there have been many attempts by political scientists to test the outcome of his thought experiment. In *Choosing Justice: An Experimental Approach to Ethical Theory* (1992), Norman Frohlich and Joe A. Oppenheimer present the results of 98 experimental approximations of the original position which they performed in the US, Canada, and Poland (1992:56).[2] Their experiments have thereafter provided a foundation for others that have wanted to test Rawls' theory.[3]

Empirical tests may seem an inconvenient method for questions of justice, the nature of which is usually thought of as strictly normative. But before we can even begin to discuss whether Rawls' conception of justice should influence how we organise society, it is crucial to assess whether his empirical assumptions are well-founded. As noted by Hare in the introductory quote: the ethics of the sorts of Rawls rely heavily on such. This is a classic trait of contractarian theories, whose truth-value ultimately is found not in the deeming of "whether such a contract has even been entered into...but whether such a contract *would* be entered into *under the specific conditions*" (Frohlich & Oppenheimer 1992:22; emphasis added).

Frohlich and Oppenheimer argued that moral philosophers are bound to end up with conclusions at odds with those of others because they limit themselves to the insufficient methods of individual deduction and introspection (1992:5). In this sense, they deem Rawls' quest to circumvent the problem of the non-existent impartial observer unsuccessful. While they agree that imperfect information is a good generator of impartiality, and that impartiality produces justice, they doubt whether the results of such can be arrived at through the mere process of *thinking*. Conclusions, they insist, must be reached empirically (1992).

The assessment that normative theories must stand testings of their empirical assumptions is an assessment which I decidedly share with Frohlich and Oppenheimer. I, however, doubt whether experiments in a lab can make *good enough* approximations to Rawls' theory to test the reliability of his conclusions. The conditions of Rawls' original position are meant to determine participants' and participants' descendants' *life chances*, which renders ethical sanctions of a well-approximated experiment highly unlikely. A laboratory would therefore be forced to significantly alter the conditions suggested by Rawls. Their conclusions can, therefore, be expected to differ from those that we would attain under ideal conditions.

My approach in this paper will be empirical; I will concentrate exclusively on Rawls' empirical assumptions about preferences under the exposure to high risk. Borrowing a distinction from Frohlich and Oppenheimer, I will concern myself with the physics, rather than with the

[2] As some focused on testing 'stability', only 70 experiments will be discussed in this paper.

[3] For examples see: R. dela Cruz-Doña and A. Martina (2000); M. Jackson and : Hill (1995); : E. Oleson (2016); G. Lissowski, T. Tyszka and W. Okrasa (1991); D. Bond and J-C. Park (1991); K. Herne and M. Suojanen (2004).

geometry, of morals (1992:24). My question is: does Rawls make valid empirical assumptions about preferences in the original position?

To start off, I will critically examine Frohlich and Oppenheimer's experiments whose results overwhelmingly point towards a mixed distributional principle which maximises the average income but imposes an income floor. I will argue that their results depend on their participants, *themselves*, attempting to become 'impartial observers' and that this hampers the impartiality-generating mechanism of the experiment. The absence of a sufficiently undesirable worst-case scenario will be argued to pose the crucial problem to their tests of Rawls' assumptions. This is because the undesirability of ending up as worst off, effectively, provides the *only* incentive for people to choose the maximin principle in the original position.[4] I will hypothesise that a sufficiently unwanted worst-case scenario is enough to neutralise the influence of people's other convictions, in which case the concealing of personal preferences is unnecessary for impartiality to be generated. To test this hypothesis I will use, as an approximation to Rawls' original position, the study of a natural experiment—*Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes* (2011), by Laura Stoker and Robert S. Erikson. Their results support that people, when exposed to a high risk, *alter* their convictions to match that which they perceive as central for the protection of their self-interests. This, I will argue, should not be interpreted as contingent on individual levels of risk aversion, but rather on a human nature—the universal desire of life preservation. Because of the radicalness of the risk associated with ending up as worst off, I will finally argue that the maximin principle is the likely result of negotiations behind the veil of ignorance. My results therefore support Rawls' empirical assumptions.

My contribution to the field of justice-related research is both methodological and substantial. On a methodological level, I criticise Frohlich and Oppenheimer's experimental approximation and, instead, suggest the use of a natural experiment. In the conclusion, I will also touch upon ways in which future researchers could make better experimental approximations of the conditions of the original position. Substantially, my results support Rawls' empirical assumptions and, arguably, also his conclusions.

[4] Inasmuch as researchers who have reused Frohlich and Oppenheimer's research design and have accepted their approximation of stakes, my criticism can be extended to cover their experiments as well.

I. Frohlich and Oppenheimer's experiments

Having identified impartiality as key to Rawls' understanding of justice, Frohlich and Oppenheimer focused on the creation of conditions of impartial reasoning which they define as "premised on setting aside one's particular interest and perspectives and giving balanced weight to the interests of all" (Frohlich & Oppenheimer 1992:3).

Experimental Design

The lab experiments were conducted with students in groups of five. They started with an educational element about the distributional principles, and then two separate stages followed (Part I and II) where monetary rewards were determined: one individually, and one collectively. The individual stage consisted of choosing the principle for four different societies, or 'situations', with varying economic realities (Table A2-A5 in the appendix). Participants were told that they could choose between the following principles[5], or, make a suggestion of their own (1992:31-51):

- Maximising of floor (*maximin*)
- Maximising the average income
- Maximising the average with a *floor* constraint
- Maximising the average with a *range* constraint

The maximum reward was $71.60[6] and the lowest was, estimably, $6.50.[7]

Results

The principle which maximised average income but with a floor constraint was found to completely dominate subjects' preferences (Table 1). Rawls' expected principle, maximin, was chosen only once (Frohlich & Oppenheimer 1992:204).

Table 1: Experimental results

	No consensus	Floor constraint	Maximin	Range constraint	Maximise Average	Total Groups
Total	**7**	**48**	**1**	**4**	**10**	**70**

[5] For Frohlich and Oppenheimer's full description of the principles, see appendix.
[6] $40 in 1992 price levels.
[7] A more thorough description of the experimental design is available in the appendix, together with motivational calculations of all estimates.

Approximation of Conditions

Imperfect information

Frohlich and Oppenheimer made a careful approximation of the veil of ignorance (1992:26). Participants remained ignorant both about the economic reality of their final 'situation' and what 'income class' they would be assigned to—how big their proportion would be (1992:27). This is to be considered a good approximation of the conditions in the original position for, as we know, Rawls' agents know neither the stage of economic development of their society, nor their particular societal position. Participants therefore had no possibility of arguing for any *certain* distribution on the basis that it would benefit them the most. What distribution benefits them most remained unclear until after the decision—when the 'situation' as well as their income class was revealed to them. Imperfect information incentivised participants to be *fair to all* and was, hence, the reason why Frohlich and Oppenheimer's thought of the resulting distributional principle as a result of impartial reasoning (1992:28).

The one part of Rawls' veil of ignorance which they were not able to approximate was, naturally, that of participants' previous values and inclinations. This will be discussed in the next section.

Agenda and motivation

In Rawls' original position, the parties are tasked with the settling of the basic terms of their association. The motivational conditions stipulate that they are self-interested and non-altruistic (Rawls 1999:102-103). Their sole concern is to further their *own* interests. The original position generates impartial reasoning *from*—and this is the whole point—*the reasoning of subjects that are completely self-interested.*

Therefore, it is surprising that we see multiple instances of suggestive wording in the experiment descriptions of Frohlich and Oppenheimer. First off, they ask their lab participants to agree to "principles of distributive justice" (1992:27). If we agree that the conditions of the original position are such as to, naturally, produce principles of fairness, then principles of distributive justice are the *inevitable result* of distribution-related decisions made by the parties, but justice should not be the 'agenda' of their discussions. Because of this, it is worrisome that Frohlich and Oppenheimer introduce the distributional principles available to their participants with the remark: "This experiment is concerned with the justice of different income distributions. Let us begin by discussing some ways of *judging the justice* of an income distribution" (1992:187; emphasis added).[8] As participants are asked to make a conscious effort to confer what justice means in relation to the distributional principles, the very intended mechanism of the original position is undermined.

[8] See appendix for full descriptions of the distributional principles.

Responding to one of their early critics, Frohlich and Oppenheimer do discuss the potential problem that the mentioning of 'justice' poses. But only in terms of participants becoming inclined to "please the experimenters" and "give *undue* and *unrealistic* considerations to matters of fairness" (Frohlich & Oppenheimer 1992:46; emphasis added). They, hence, discuss this as if it were a problem of external validity; an inevitable problem of the experimental setting. They do not problematise the fact that discussions about justice are not—at all!—the *supposed* agenda of parties in the original position. The below excerpts from the experiment transcripts inform us that this was a problem. As is evident, when participants discuss how to deal with the 'worst off' in society, they do so by referring to 'them', in third person, and not to a potential 'I' or 'we'. Self-interest (expressed through first-person references) seemingly only influenced participants when discussing the position of more privileged citizens; a position which they, presumably, could identify with personally (they were all students at institutions of higher education). When addressing the 'issue' of 'poor people', they distanced themselves from this position and, *themselves,* attempted to incorporate the virtues of an impartial agent (1992; emphasises added throughout the quotes):

"If you have people that are really really poor, … *they* have a tendency to just stay there (…). But (…) [with] a certain minimum, then *they* have a chance to get out of that situation." (61)

"...[if the floor is too low]… *people* are going to be starving, and *they* will all be without shelter and housing" (61)

"I suggest we choose the thing with the largest range of distributions. That way at least somebody is going to get a real good payoff." (62)

"I'm actually thinking of the *poor people*—like, I don't want to see people starve to death, but I don't want anybody to limit *my income* just because it's some sort of socially adopted policy." (63)

"I think we definitely need a floor constraint… in terms of justice, it's fair". (63)

"I'm not looking at it from an individual point of view. I'm looking at it as if I'm some kind of god. I'm looking down on what is best for society, not what is best for me… I think the thing is to maximise the average." (111)

Frohlich and Oppenheimer do, to be fair, comment on the last of the quotes by saying that "most discussions went smoothly without imaginary roles of god and the like" (1992:112), signalling that this was not how they intended for participants to reason. But they do not seem to find problematic the way that the rest portray the focus of the discussions. Instead, they specifically express contentment when noting that "subjects seemed seriously intent on discussing need, entitlement, and incentives, and the appropriate tradeoff between them" (1992:162). The reason why Rawls expects parties to support the maximin principle, however,

lies in his deeming of it as the most suitable for anyone aiming to protect their *own* self-interest—not because they should be "moved by the ethical propriety of the idea" (Rawls 1999:156). Frohlich and Oppenheimer demonstrably overlooked the essential mechanism of the original position, and this critically undermines their study.

Stakes

The stakes faced by parties in Rawls' original position are extremely high. Their entire life chances (and those of their descendants) depend on the combination of natural talents and social position they end up with (Frohlich & Oppenheimer 1992:27). The stakes involved in Frohlich and Oppenheimer's experiments can best be formulated as the opportunity cost associated with leaving the experiment with a smaller reward than one would have wanted. The biggest reward was $71.60 and the smallest was, estimably, $6.50 (Frohlich & Oppenheimer 1992:45).[9] Needless to say, stakes were much smaller.

I argue that impartiality, while it can be *induced* through restrictions on information, is *ensured* through the presence of high and personal stakes. Unless they are high, we lack incentives to adapt our reasoning to the conditions stipulated by the situation at hand and may, instead, make judgements based on our personal convictions. In the real world, high stakes are imposed to ensure that *due* weight is attributed to all considerations.

Consider the case of a judge. Consider that we, after his or her ruling in a case, find that due weight has not been given to all sides, not because the judge was personally involved in the issue at hand, but because the version offered by one of the sides was more in tune with the judge's own life philosophy and was easier for the judge to empathise with. The case may have been related to something which the judge felt strongly about: abortion rights or gun laws. The judge would then risk legal and even punitive repercussions that have been put in place with the sole purpose of ensuring that the judge bases his or her ruling—exclusively—on the conditions provided by the court. This shows that the 'mere' risk of having one's ruling dissolved (and one's reputation damaged), has not been deemed to provide *high enough* stakes to ensure a credible level of impartiality. Clearly, high enough stakes are usually considered important for judicial impartiality to be credible. Let us consider another example.

Excessive risk-taking by banks has been shown to have consequences as far-reaching as the disruption of entire markets—affecting countries and people on all levels. Safe to say, stakes are extremely high. The case has, however, been made that bankers have not *felt* the level of risk at hand, since governments often 'bail out' banks in trouble, essentially reducing the stakes through covering for them financially. This is because societal costs of banks failing are thought to be higher than the costs associated with bailing them out. In their article *The Impact of Public*

[9] See Appendix for calculations of estimates.

Guarantees on Bank Risk-Taking: Evidence from a Natural Experiment, Gropp et al. (2014) analyse the impacts of the removal of explicit federal guarantees on a few german banks' risk-taking behaviour. They conclude that banks, when government guarantees were removed, reduced their exposure to risk in virtually any way they could (2014:457). While bankers are not expected to act with the same desired level of impartiality as a judge, their behaviour shows something important about the immediate effect of accountability on people's risk assessments. When stakes are high and personal, risk-taking behaviour decreases.

Returning to the experimental approximation of Rawls' original position, I argue that stakes neither were high nor personal *enough* to ensure that participants base their judgments on potential outcomes—at least not in the way we may expect them to in situations characterised by the same levels of risk as the original position. Instead, participants probably either supported the principle which was most likely to maximise their expected payoff (like the bankers), *or* the distributional principle which best matched their personal convictions (like the case with the judge).

In the case that the stakes are so low that the worst-case scenario turns into one which *really is not so bad*, it may seem unreasonable to lower the expected payoff only to protect oneself from ending up there. In the case of the experiments in lab setting, the worst-case scenario was equivalent to leaving the experiment with 'only' a small reward: $6.50. To receive this little, you had to

1) be so unlucky as to end up in the lowest income class under the application of the distributional principle which minimised earnings for the lowest income class
2) in each of the *four* educational 'situations' in part I of the experiment,
3) *and* in part II.

Conversely, the worst-case scenario if one, repeatedly, chose the maximin principle was $9.90. Leaving with a $6.50 reward is, arguably, not *that* different from $9.90. It would then seem more rational to support an option which, while lowering the reward for the *worst* case scenario (by $3.40), greatly increased their expected payoff (average income). This way, one would at least have secured the *potential* for a significant reward if one was lucky enough to end up in a better position after the random assignment of income classes. If worse came to worst, one would leave the experiment $6.50 richer, a scenario from which people are unlikely to have gone to great lengths to *protect* themselves from—especially not when that protection, while not guaranteeing a considerably better outcome, came at the cost of lowering their expected payoff.

When the worst-case scenario is not associated with a significant risk it seems odd that participants should pay much attention to a principle the strength of which entirely lies in its ability to protect them from the negative effects associated with ending up in the least desirable position. Strictly speaking, the absence of significantly negative potential consequences has the

effect of removing virtually any incentive to support a 'maximin' distributional principle. This should mean that experimental approximations that fail to present significant risks to its participants have very limited explanatory power concerning Rawls' original position.

Table 2: Total Payoffs Of Part I+II* When Best Off / Worst Off / Average Income, Per Principle (2018 Price Levels) (Percentage Of Max Income In Red)

	Best off**	Average income***	Worst off****
Maximisation of average income	$62.80	$32.3 (45,1%)	$6.50
Maximisation with floor constraint	$61.50	$28.9 (40,4%)	$8.95
Maximisation with range constraint	$55.80	$25.9 (36,2%)	$8.70
Maximin	$69.30	$25.6 (35,7%)	$9.90

* All totals are calculated with assumption that the same principle had been chosen repeatedly

** Assumption of max payoff in part II = $33.50, for all principles

*** Assumption: average income in part II had the same relative size to the max income as it had in part I

**** Assumption of min payoff in part II = $0.50, for all principles

Now, if it is clear that it is not 'rational' to protect oneself from the worst-off position when stakes are low, how come participants so clearly preferred the principle with a *floor constraint?* In Table 2, we can see the effects of repeatedly ending up as *best off*, *worst off*, or with the *average income*; in both part I and II, and in the case that any of the distributional principles has been chosen repeatedly.[10] Following my line of argument, we should expect participants to be more concerned with the maximisation of expected payoff than with the amelioration of the worst-case scenario. The stakes, I argued, were *too low* to induce such a concern. The difference between winning $6.50 and $8.95 (the minimum payoff under the principle with the floor constraint) is, however, not so big either. I argued that the $3.40 difference between (repeatedly) ending up as worst off under the principle which maximises average income compared with the 'maximin' was not *big enough* to render 'maximin' desirable. But by preferring the floor constraint principle over that which maximises the average they sacrificed $3.40 of their expected income *only to secure the guarantee of $2.45* in the case they ended up as worst off. Why not just go for the maximisation of average income? As it turns out, the choice of

[10] All of the assumptions are accounted for in the appendix.

principle had a very marginal effect on participants' payoff, which instead overwhelmingly was determined by the income class that they, randomly, were assigned to. What Frohlich and Oppenheimer's experimental stakes, hence, entailed was that neither distributional principle could do much to improve prospects in terms of monetary payoff.

It is, indeed, surprising that these small differences should have been able to produce such clear and unanimous support for any of the principles at hand.[11] This may mean that participants preferred the principle with the floor constraint because it was most in accordance with their personal preferences. This would go hand in hand with the reasoning of Chong et al, who in their article *When self-interest matters* (2001) argue that people shape preferences according to their values instead of self-interests when benefits are not large or clear—"especially when exposed to information that cues sociotropic concerns, group identifications, or value-orientations" (p:544-545). As we recall, participants were indeed cued to think about justice as they decided on the distributional principles. Since decisions probably were made on the basis of participants, themselves, attempting to reason impartially (rather than self-interests *created by the conditions of the experiment*), I suggest that the original position created by Frohlich and Oppenheimer was unsuccessful in generating impartiality.

Experimental Variations

Some experiments were made with alterations to the design. As is visible in Table 3, where we can see the results of all the different experimental variations, the principle with a floor constraint was the winner regardless of experimental variation (1992:204). We can, nonetheless, see that the alterations did have a significant effect on the outcome.

Six experiments were conducted where all references to 'justice' were removed, to test for biases caused by a wish to "please experimenters" (1992:46). Out of six samples with this variation, four were found to result in the principle with a floor constraint, and two resulted in the maximisation of the average income. While the alteration did not produce results which were more in line with Rawls' expectations, they did change. Maximisation of average income, which in the other experiments had been chosen only 8 times out of 64, were in these experiments chosen in as many as two out of six cases. This means that the ratio of preference for the principle which maximised the average more than doubled with this variation—going from 1/8 to 1/3. This supports the suspicion that the original position's impartiality-generating

[11] Participants may not have gone through the effort of calculating as thorough comparisons between their expected payoffs in each of the income classes as I have. In each of the 'situations' in part I, however, when they had randomly been assigned with an income class, they were also provided with a chit (see Table A1 in appendix) with information about what they would have gotten if they had chosen any of the other principles. These were later kept by participants to enable them to look back on previous consequences of their choices, and Frohlich and Oppenheimer assess that this had a great pedagogic impact on participants (1992:38). Comparisons of this sort are therefore not unlikely to have happened.

mechanism was distorted as participants, themselves, attempted to act impartially. This may not necessarily be, as was conjectured by Frohlich and Oppenheimer, because they wanted to "please the experimenters" and "give undue and unrealistic considerations to matters of fairness" (1992:46), but simply because they were instructed to do so. Unfortunately, the small number of experiments conducted in the alternative way prevents us from drawing any definitive conclusions, and it is therefore regrettable that Frohlich and Oppenheimer chose not to conduct more than six experiments of that sort.

TABLE 3: RESULTS BY EXPERIMENT TYPE

(Frohlich & Oppenheimer 1992:204)

	No consensus	Floor constraint	Maximin	Range constraint	Maximise Average	Total Groups
Regular stakes with gain	7	23	1	2	1	34
'Nonjustice'	0	4	0	0	2	6
Regular stakes with loss	0	11	0	0	5	16
High stakes with gain	0	6	0	1	1	8
High stakes with loss	0	4	0	1	1	6
Total	7	48	1	4	10	70

Two experimental variations were made to test participants' sensitivity to higher stakes: 'higher-variance payoffs' and 'losses'. In the first of these, floors were made lower and ceilings higher. Frohlich and Oppenheimer expected that if Rawls was right in suggesting that "rational individuals consider nothing but the floor when selecting a principle of justice", this would have the effect of making the worst off-position even less desirable, and therefore the maximin principle might appear more attractive (1992:44). If ceilings, simultaneously, significantly went up, however, one could also conjecture that people would find more attractive the potential of ending up as *best off*, and followingly support principles with the better payoff potentials.

The variation which involved 'losses' was introduced since it had been put forth that people react with greater risk aversion in situations that can lead to losses rather than gains. In these variations, participants were given a $71.60 credit sheet at the beginning of the experiment. Their payoff was subsequently determined through the subtraction of rewards instead of additions (1992:45). This way, they may have experienced the situation as one where they had to protect 'what was already theirs'.

With the introduction of higher variances and perceived losses, the principle which maximised average income, again, gained ground. In the standard experiments, with regular stakes and 'gains', maximisation of average was selected only 1/34 times. With the alterations of the stakes, it was selected 5/16 , 1/8, and 1/6 times, respectively. As stakes (real or perceived) grew, participants hence became more likely to support the principle which maximised their expected payoff: average income. These results are more in line with my previous reasoning. Although stakes clearly began to get interesting enough to make participants rely on self-interest instead of personal values (remember Chong et al), they still lacked a truly undesirable 'worst off'-position from which it would have been rational to try to protect themselves. In order for the approximation to be convincing, I argue that it would have been necessary to introduce higher stakes and—especially—worse 'worst off'-positions.

EVALUATION

Is Frohlich and Oppenheimer's approximation of of Rawls' original position *good enough* to test what principles would be chosen by parties behind the veil of ignorance? Based on my examination of the approximated conditions, one could criticise the validity of the experiments because of 1) excessive influences from participants' own convictions 2) the absence of a sufficiently undesirable possible outcome.

As I have argued, it seems likely that participants chose to maximise the average *with a floor constraint* because it best resonated with them when they, themselves, attempted to reason as impartial agents. Interest for simply maximising the average significantly increased when they were presented with alterations in the experimental design, both as references to 'justice' were removed and when stakes (real and perceived) were raised. Since the greatest risk faced by participants, regardless of the experimental type, still involved *receiving* a monetary reward, the experiment cannot be said to have tested participants' propensity to protect themselves from a worst-case scenario worthy of the name.

In order for my criticism to be convincing it would, however, not be enough that I show that the *experimental results* largely depended on either (or both) factor 1) and 2). I would also have to argue that these factors are crucial for the determination of outcomes in Rawls' original position. In the next section, I will argue that the exposure to a risk of a sufficiently undesirable outcome would have been enough, both to neutralise the influence of previous preferences, and to increase support for the maximin principle.

II. The Relative Influence of Self-knowledge and Risk Under the Veil of Ignorance

Self-knowledge

While successfully implementing the veil of ignorance in terms of income classes, Frohlich and Oppenheimer's experiments were, understandably, unable to conceal participants' *self-knowledge:* about their own personalities, values, and 'conceptions of the good'. As I argued at length in the previous sections, this may lead us to mistrust the reliability of their so-called 'generated impartiality'.

However, Yale professor Stephen Darwall argues that Rawls' harsh self-knowledge restriction, *in itself*, poses a problem for the value of his original position (Frohlich & Oppenheimer 1992:17-18). If principles which result from negotiations behind the veil of ignorance appear desirable only in the *complete absence* of values and preferences, the use of the resulting principles may be questioned. Humans are full of values and preferences. As these values might even be what makes us human in the first place, it may seem absurd to think that we could arrive at just principles only through the *elimination* of the influence of values. This might just have the effect of making us so dehumanised as to render choices made under such conditions irrelevant for generalisations about human behaviour. A counterargument is that the removal of personal values could reveal something which truly underscores our humanity, since what is left at that point—if anything at all—represents something we all have in common. This stance would call for notions of a certain human nature.

It is, of course, impossible to find out what happens if we remove all personal characteristics and preferences from a person. But it might be that certain conditions lead people to *disregard* personal preferences, in which case they could be deemed irrelevant for the outcome. This would mean that we, as long as we are able to neutralise the influence of other convictions, may be able to produce a valid setting for impartial reasoning—*despite* the obvious impossibility of recreating Rawls' ideal condition of concealed self-knowledge. It may even be so, that the stakes involved in the original position are examples of such conditions. This, in turn, would entail that concealed self-knowledge is an unnecessary trait of the original position for impartiality to be generated. This brings us to my next criticism of Frohlich and Oppenheimer's approximations.

Risk

Despite recognising the obvious discrepancy between stakes determining 'life chances and those of your descendants' and ones that 'guarantee you with a reward ranging from $6.50 to $71.60', Frohlich and Oppenheimer "doubt that the results of experiments [with higher stakes]

would be very different" (1992:162). They base this assessment on the "tone of the discussions", and conclude that the attraction of the floor-constraint principle lies in its enabling of tradeoffs between competing values of interest (1992:30):

- needs of individuals unable to care for themselves
- entitlements to the fruits of their labour
- the economic need of incentives for productivity
- the trade-offs among the floor, the ceiling, and the mean

Followingly, they expect that agents in the original position, while likely to care about the "aspect of the welfare of the poor", will not ignore "the rewards of those who work hard and are productive" (1992:30).

Why, then, is Rawls so convinced that agents in the original position would be willing to sacrifice the prospects of these rewards? The results of Frohlich and Oppenheimer's experiments decidedly indicated that these rewards matter (perhaps even the most—if we consider the trend of the variations). Rawls assumes people to be willing to sacrifice even large parts of their expected utility to secure their own self-interest through the guarantee of a decent—in relation to the rest of society—standard of living. As it turns out, it is not the sacrifice of 'rewards of the productive' which Rawls considers most worthy of consideration. He identifies a competing sacrifice, the reluctance to which he thinks will trump that of expected utility:

> "The principles of justice apply to the basic structure of the social system and to the determination of life prospects. What the principle of utility asks is precisely a sacrifice of these life prospects. Even when we are less fortunate, we are to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of a life. This is surely an extreme demand. In fact, when society is conceived as a system of cooperation designed to advance the good of its members, it seems quite incredible that some citizens should be expected, on the basis of political principles, to accept still lower prospects of life for the sake of others." (1999:155)

A counter argument could be that the sacrifice of rewards for productivity *also* would be regulated by the basic structure of the social system and, hence, *also* would apply to the determination of entire life prospects. Why would we, then, prefer to sacrifice higher productivity rewards over sacrificing the guarantee of higher life prospects?

Under the maximin principle, inequalities are only accepted if the inequalities, themselves, are to the benefit of the worst off. Formulated in terms of agents that end up as 'better off' under

the application of a *maximin* principle, the sacrifice would entail that the rewards for their productivity, over the entire course of their lives, would be constrained by the binding link to the situation of the 'worst offs'. The suggestion is that the 'better offs' would accept this since they, themselves, would have opted for this option in the original position, before knowing which position they were to end up with.

It does not come as a surprise that many should find this uncomfortable. On the contrary: in discussions about redistribution in society, the absurdity of forced benevolence is usually invoked, and the justifiability of this institutionalised sacrifice is therefore, more often than not, questioned.

Formulated in terms of agents ending up as 'worst off' under the application of a principle which *maximises the average income*, the sacrifice entails the acceptance of lifelong *lower* life prospects compared to a society organised under the maximin principle. The relative effects of ending up as 'better off' are widely known: longer life expectancy, better health, access to health care, more free time, room for failure without devastating consequences, opportunities to travel, the privilege of being taken seriously, social trust, etc. These effects are likely to apply regardless of the economic stage of development of the country in question.

The acceptance of lower life prospects would be justified with the assertion of greater importance of high productivity rewards for those who *already* were lucky enough to end up as better off. Again, the 'worst offs' would be expected to accept this situation, since they *themselves,* would have preferred this way of societal organisation in the original position, before they knew whether they would end up as better or worse off.

Seeing how we do not live in societies where basic regulative principles have been determined through the agreement by impartial parties in a contract, the so-called 'sacrifice of people ending up as 'worst off' is, for the most part, not thought of as a sacrifice at all. It has never been asked to be tolerated by the subjects themselves, and, therefore, it is not one which requires justification.[12] A sacrifice is, nonetheless, the only way to regard consequences of ending up as worst off when we address them through the prospects of agents choosing principles behind a veil of ignorance. A principle which asks people who *already* have been assigned with an undesirable position to keep accepting *even lower* life prospects for the enabling of higher life prospects of those who are *more* fortunate would, from the perspective where agents have been presented with a choice, also undoubtedly be characterised by its fair share of 'benevolence' (Rawls 1999:157).

[12] Climate activist Greta Thunberg did, however, make such references in her speech at COP24: "Our civilisation is being sacrificed for the opportunity of a very small number of people to continue making enormous amounts of money. Our biosphere is being sacrificed so that rich people in countries like mine can live in luxury. It is the sufferings of the many which pay for the luxuries of the few" (CNN, December 16 2018).

The question at hand is what sacrifice people will be most reluctant to make; the sacrifice of productivity rewards for the welfare of the worst off, as one ends up as better off under maximin; or the sacrifice of higher life prospects for the higher rewards of the better off, as one ends up as worst off under maximisation of average? While 'annoying' might be an adequate description of the sacrifice of the 'better offs' under the maximin principle, Rawls believes us to find the sacrifice of the 'worst offs' under the maximisation of average utility "intolerable" (1999:155).

Frohlich and Oppenheimer claim participants' differing levels of risk aversion to lead to the preference of the mixed principle, since this is considered to represent the middle ground. This interpretation coincides with the conclusions drawn by Roger Howe and John Roemer, who, when modelling the original position as a game, framed variations in preferences as a function of agents' aversion to risk. They therefore assess Rawls to *assume* high aversion to risk (1992:28-29).

Rawls would, however, not accept this assessment: "The principles chosen [must not] depend on special attitudes to risk. For this reason the veil of ignorance also rules out the knowledge of these inclinations" (1999:149). Considering how huge of an incentive (if not the only!) the risk of ending up as worst off (under the principle which maximises average utility) is for the maximin principle to be preferred, it seems ridiculous not to acknowledge the role of risk aversion in deeming this the rational solution. What Rawls means, however, is that one's *individual* level of aversion to risk will not play a big role once we find ourselves behind the veil of ignorance and this, he argues, has to to with the *unique features* of the situation. These features are what makes the maximin principle the only rational option *"*for anyone *whose aversion to uncertainty* in regards to being able to secure their fundamental interests *is within the normal range"* (1999:149; emphasis added).

This, nonetheless, clearly indicates that he *does* assume a 'universal' aversion towards risk, and this may be hard to swallow. One may, just as well, interpret this not as 'ruling out' such knowledge, as Rawls puts it, but rather as an assignment of a 'certain level' of aversion to risk. A suspicion may be raised that Rawls not only tries to disguise his assumption of a 'certain' level of risk aversion as 'neutral' by claiming that we have no knowledge of how big it is, but that this is also one he greatly depends on in order to arrive at the conclusions which he, himself, prefers. The assumption should not be disguised as 'neutrality', but ought to be argued for. The assumption is that when the risks are undesirable *enough*, one's 'particular' level of risk aversion becomes irrelevant, since some risks are so great that they have the effect of inducing high levels of risk aversion in *anyone*. The higher the risk, the greater the aversion.

Tellingly, Rawls did not even think that the stakes could be described as a game where you can gamble for money (Wolff 2010). In such a game, risk aversion would surely play a great role. However, seeing how 'risk aversion' varies between individuals, I agree with Rawls that it may not be an appropriate term to describe the mechanism which he assumes in the original position. A universal trait should, instead, be treated as part of human nature.

In his justification for state sovereignty, Rawls' liberal and contractualist predecessor Thomas Hobbes (1651) argued that it was rational for people to give up *all* their freedoms in exchange for physical protection (Rosen and Wolff 1999:58). Just like Rawls, Hobbes imagines an original position or a 'state of nature' where he limits his description of the characters to their one and only desire in common—that of life preservation. This, he believes to have priority over all other, *however differing*, desires (Moehler 2018:42). While Rawls makes sure to establish a priority of the protection of the freedoms that Hobbes sacrifices, he still seems to believe in the feasibility of negotiating a better deal concerning the guarantees that the sacrifices are swapped for. In his case, the guarantee of life quality extends beyond the mere physical protection of the person, and so he manages to involve a decent level of material welfare (relative to that of the rest of society)—all to capacitate individuals to lead a life according to their own convictions of what such a life consists in.

Hobbes did not think it necessary to motivate the sacrifice of all personal freedoms with promises of milk and honey. He deemed sufficient the threat of having to live a life without a guarantee of physical protection for it to be rational for men to accept the sacrifice, and he thought so because of the radicalness of that threat. He famously described such a life as characterised by "continual fear, and danger of violent death; and the life of man: solitary, poor, nasty, brutish, and short" (Rosen and Wolff 1999:13). Thus, while their views differed in many ways, the conviction of a common, life-preserving, interest being part of human nature is something which Rawls shared with Hobbes.

Frohlich and Oppenheimer are, nonetheless, unlikely to be impressed by the stressing of further theoretical conjectures about matters which really require empirical analysis for reliable conclusions to be drawn. Their frustration with the deductive and introspective traditions was, as we recall, precisely what made them embark on their empirical investigation of justice in the first place (1992:19-20). Luckily, we are not forced to rely on 17th century philosophers' ideas about human nature to test the extent to which people are willing to overlook other—competing—inclinations when their capacity to secure their fundamental interests is seriously threatened. Instead, we are better off consulting endeavours similar to those of Frohlich and Oppenheimer. In what remains of the paper, I will empirically examine the robustness of the theoretical reasoning of the sorts of Hobbes and Rawls. Through an approximation of the

conditions of the original position to a natural experiment, I will test the sensitivity of preferences to the exposure of high risk.

III. Vietnam Lottery Approximation

"In my judgment, a fair system is one which randomises by lot the the order of selection. Each person in the prime age group should have the same chance of appearing at the top of the draft list, at the bottom, or somewhere in the middle. I would therefore establish the following procedure."

Richard Nixon (May 13 1969:4)

The late sixties in the United States were marked by protests from the draft-resistance movement. In Vietnam the war was escalating, and as more men were getting called in to serve the arbitrary distribution of exemptions between draft boards across the country became increasingly evident (Stoker & Erikson 2011:222). Because of the shortcomings of the old draft system, a new system was implemented in the form of a lottery where birthdates would be drawn to randomly decide on an order in which men would be drafted. This meant the end of many categories for exemption. The lottery draw was broadcasted on television on December 1st 1969 and the first birth date to be drawn, and hence to be called under the new system, was September 14th (2011:222). It was unclear how long the war would go on for and, therefore, how many men would have to be drafted.

In their study *Caught in the Draft: The Effects of Vietnam Lottery Status on Political Attitudes* (2011), Laura Stoker and Robert S. Erikson use data from surveys and interviews with high school seniors from the class of 1965, before and after the national draft, to see how the exposure to the high and uncertain risk of getting drafted affected their political preferences. Interviews and surveys were conducted in 1965—before subjects knew that there was even going to be a lottery—and in 1973, when their draft related faiths were already known to them. The subjects of interest for the study had all, prior to the lottery, been exempted from the military draft thanks to their college-bound status. The implementation of the lottery system meant the end of this category of exemption, and this is why their faiths were very clearly affected. This was not the case for men who were not college-bound; they had already faced the risk of getting drafted *before* the implementation of the lottery. These men are, therefore, used as a control group (Stoker & Erikson 2011:224-225). In 1973, the US was just about to call their troops back and it had been known since 1970 that the cutoff number was going to be 195 (2011:225). As we will see in the results, however, the effects of draft vulnerability were still highly present. The surveys included questions that either related to attitudes towards the Vietnam war or towards specific political issues; candidates; party identification; or ideology.

Approximation to the original position

The logic behind the approximation of their study to Rawls' original position can be found in the investigation of effects on attitudes of the exposure to high but uncertain risk. Rawls judges these particular features of the original position to be what will lead to the preference of a maximin principle of distribution. Previous tests have, as I have argued, been unable to evaluate the effects of a sufficiently *undesirable* worst-case scenario which is the crucial incentive to support the maximin principle. Judging from the introductory quote by Nixon, he deemed the worst-case scenario of the draft to be such as to require a drafting procedure characterised by fairness and, in this case, his idea of what this entailed was curiously similar to that of Rawls.[13]

The subjects of Stoker and Erikson's study were exposed, from one day to another, to a situation characterised by a sufficiently undesirable worst-case scenario to resemble that imagined by Rawls. Having access to data of their political preferences before and after the exposure to this risk, Stoker and Erikson have been able to assess the effects of such exposure. The results of the approximation will, hence, be discussed in terms of how subjects' attitudes changed as a consequence of being exposed to the risk of getting drafted.

By testing the influence of exposure to high risk on attitudes, I will test my own criticism of the lab experiments. If attitudes should remain unchanged despite the exposure to high risk, Frohlich and Oppenheimer would have made a correct assessment, deeming preferences to be insensitive to increases in stakes. This would, in turn, suggest that *imperfect information* is the crucial explanatory determinant of outcomes in Rawls' original position—not *high stakes*, as I have argued. We would, in that case, have to trust the production of imperfect information to provide sufficient basis for the generation of impartial reasoning. Results will, however, decidedly suggest this not to be the case.

Conditions

Imperfect information

We are, in both the case of the original position and in that of the Vietnam Lottery, interested in *potential*, very radical, changes in self-interest. One can distinguish between self-interests that are contingent on circumstances that are certain, on the one hand, and on circumstances that are not certain, on the other. In both our cases, we have to deal with interests that are contingent on circumstances that are *not* certain; information is, in other words, imperfect. The exposure to this potential outcome has no connection to one's previous life situation; the outcome cannot be explained by factors that relate to one's previous circumstances or choices,

[13] Given that the new draft system was implemented by the Nixon administration in 1969, and that Rawls, prior to publishing *A Theory of Justice* in 1971 had written academic articles about Justice for more than ten years (from prestigious academic institutions such as Princeton, Oxford, Cornell and Harvard), it does not seem unlikely that the new draft system, itself, contained immediate influences from Rawls' philosophical ideas about fairness.

but is rather the consequence of random selection. In the original position, this is ensured by the random assignment of societal positions (and natural talents). In the case of Stoker and Erikson's study, it was ensured through the random assignment of lottery numbers which were connected to subjects' birth dates—something which is not considered to have a systematic effect on people and which, subsequently, is why we can treat their differences in attitudes as direct effects of draft vulnerability.[14]

Because of important differences in the *levels of vulnerability* faced by men in the Vietnam lottery study compared to agents in the original position, I will first clarify which lottery numbers I deem suitable for an approximation.

Agents behind the veil of ignorance are *equally vulnerable* to a potential and very damaging outcome (henceforth *PVDO*); no one is more prone to the risk of ending up in a bottom position than another. Likewise, the college bound men were *equally* vulnerable, before the lottery, to the *PVDO* of receiving a lottery number which could lead to them getting drafted in the Vietnam war. In an ideal scenario, we may have had access to a measure of the political attitudes of the young men when they had been informed that there was going to be a lottery but before they knew what number they were going to be assigned with. This could have been said to ensure information about the political attitudes of the college bound men when they, just like the agents behind the veil of ignorance, were *equally vulnerable* to the *PVDO*. The consequence of the random selection of the lottery is, however, not exactly the same as the consequence of lifting the veil of ignorance.

While the lifting of the veil of ignorance leads directly to the *outcome* to which the parties previously have been vulnerable (in the approximated case, the actual drafting of specific men), the random selection process represented by the lottery draw only opens up to a *new moment of vulnerability* (the period where the lottery results have been revealed but the actual drafting is yet to happen). The *lower* the lottery number that one was assigned with, the *higher* was the risk of getting drafted, and so, in the period following the announcing of the lottery outcomes, the relative extent of vulnerability differed between the subjects. This means that they now, as opposed to the parties in the original position, were *unequally* vulnerable to the *PVDO*. So, while the veil of ignorance presents a situation which can be described through the following scheme,

[14] It should be noted that it was discovered that there had been a systematic bias of lottery numbers coming from the last two months of the year as a result of these months having been put into the box last and the box not having been shaken sufficiently. As a consequence, men born in November and December faced a higher risk of receiving a low lottery number than did men born in the earlier months of the year. When analysing the effects of this bias statistically, however, Adam J Berinsky and Sara Chatfield (2015:453) concluded that "researchers can largely treat the lottery assignment as if it were random".

1) Equally vulnerable to *PVDO* —> 2) outcome (social status and natural talents)

the situation of the college bound men that were subject to the Vietnam lottery should rather be described as follows:

1a) Equally vulnerable to *PVDO* —> 1b) *unequally* vulnerable to PVDO —> 2) outcome (drafted or not)

As suggested by the scheme, then, in the case of the Vietnam lottery we have *two moments* of vulnerability (1a and 1b), as opposed to the ideal of only one, and in one of the moments in the Vietnam lottery situation (1b) subjects' relative extents of vulnerability *differ*. This is not ideal for the approximation. We may, however, make an approximation by focusing on only one of these moments. Since we lack data of political attitudes of subjects in moment 1a (when the college bound men had found out about the lottery but had not yet been assigned with a number), we are, regardless of preference, forced to focus only on 1b. The situation of the subjects in 1b may be considered better suited anyway, since it is the one 'nearest' the outcome (in time), and therefore similar to the situation of the parties under the veil of ignorance. But the subjects' unequal exposure to vulnerability causes problems for the approximation, and we therefore have to, somehow, categorise the subjects into groups characterised by similar vulnerability levels. Here, we first need to clarify whether the *undesirability of the outcome* or the *uncertainty* should be considered the critical constituent of vulnerability. *Uncertainty*, in the case of the subjects to the Vietnam lottery, refers to ignorance about whether one will end up getting drafted or not. *Undesirability*, on the other hand, refers to the peril involved in getting drafted.

As for the differing levels of uncertainty, numbers can be categorised as either 'low' (1-122), 'middle' (123-244), or 'high' (245-366), where low and and high numbers all face lower levels of uncertainty (they are pretty sure that they either *will* or *will not* be drafted), whereas subjects with middle numbers face high levels of uncertainty (it depended on how long the war would go on for). As for the level of undesirability faced by the college bound men, it had an inverse relationship to the progression of lottery numbers: the higher the number, the lower the level of risk of the undesirable situation.

As parties behind the veil of ignorance confront a situation with very high levels of both undesirability and uncertainty, this would, ideally, also be the case for our men in our chosen category of lottery numbers. Unfortunately, as is evidenced below (Figure 1), this ideal cannot be reached with our sample.

FIGURE 1: VULNERABILITY LINE

Number (1-366):	low (1-122)	medium (123-244)	high (245-366)
Uncertainty:	low	high	low
Undesirability:	high	medium	low

It could be argued that either the lowest numbers or the middle numbers represent the most useful group for an approximation, because they are the alternatives which either face high risk of the undesirable outcome *or* high uncertainty. This kind of differentiation in numbers is, unfortunately, not possible with our data. Stoker and Erikson created an index from 0-1 of the numbers 1-366, and their results are therefore all expressed as the expected change in attitudes, going from the holder of lottery number 1 to lottery number 366 (2011:226).

TABLE 4: MEDIATING ROLE OF MILITARY SERVICE

(Stoker and Erikson 2011, p.228)

	Dependent Variable = Composite Vietnam War Attitude, 1973		
	OLS		
	All College Bound ($n = 255$)	No Military Service ($n = 172$)	Military Service ($n = 83$)
Lottery number	0.24 (0.08) $p = 0.002$	0.30 (0.09) $p = 0.002$	0.10 (0.12) $p = 0.396$

Notes: All results for college-bound (those whose 1965 high school curriculum was college preparatory) males who did not enter military service prior to 1969. Standard errors are clustered standard errors. All variables are scaled 0 to 1.
[a] Instrument for Military Service is a 0-to-1 dichotomy, whether the draft number was 196 and above or 195 and below.

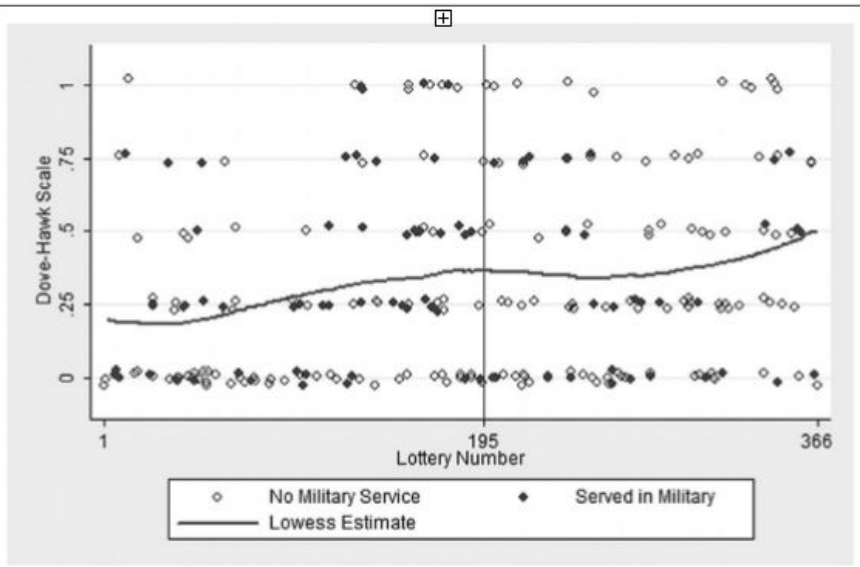
We do, however, have three good reasons to expect a high level of uncertainty *even* among holders of lower numbers (1-122), as well as for significant changes in attitudes to result *even* in the case where we, instead of indexing numbers from 1-366, were able to index them from 61-366 (which would be the ideal for our approximation).

1) Only 32% of the sample, 39% of holders of lottery numbers below 195 and 24% of those above, ended up serving (Stoker & Erikson 2011:225). The fact that over 60% of holders of numbers below the cutoff (195), in the sample, ended up *never serving* indicates that they impossibly could have been *certain* about having to get drafted. Uncertainty amongst those with lower numbers can therefore not be ruled out.
2) Table 4 presents the expected change in war attitudes going from the lowest lottery number to the highest, with war attitudes indexed from 0 (dovish) to 1 (hawkish). When looking at

the entire sample, we can see that attitudes became 0.24 increments more hawkish between the holder of the first number and that of the 366th. As 'no military service' is held constant (third column), testing the potential mediating role of military service, the expected change increases to 0.30. This means that men who ended up *not serving,* and amongst whom we can expect a lower frequency of holders of the very lowest lottery numbers (and that we prefer not to look at since they are 'too certain' for our approximation), showed greater variation in their attitudes towards the war than was shown when *all* were considered. This suggests that men with low-to-medium numbers that ended up *not* getting drafted, and who, because of this, clearly *cannot* have been 'certain' that they would get drafted, *still* adopted attitudes towards the war that were greatly influenced by the vulnerability caused by their lottery status. As it seems, the expected difference in attitudes between holders of the lowest and the highest number, amongst those *who were never even forced to serve,* is even greater than that which is expected when *also* considering those who served (0.3 compared to 0.24).

FIGURE 2: LOWESS CURVE DISPLAYING WAR ATTITUDE AS A FUNCTION OF DRAFT LOTTERY NUMBER

(STOKER AND ERIKSON 2011, P.227)



Notes: Observations are slightly jittered in order to enhance the visual display. The LOWESS curve predicts the Dove–Hawk score based on lottery number for all observations in the graph, regardless of actual military service. Bandwidth = 0.8.

3) Figure 2 indicates that the adoption of more 'dovish' war attitudes was not limited to the 'unsuitable' group of lottery numbers. The group which I argued was better suited, those with numbers ranging from 61-184, adopted attitudes to the war that estimably range between 0.23-0.32 on the 0-1 scale, compared to those who, in terms of lottery number, were completely 'safe' and that averaged around 0.5. The effects of being vulnerable to the draft are, hence, not limited only to numbers between 1-60.

I argue that, although not ideal, Stoker and Erikson's results *can* be discussed in terms of the effects of an *uncertain* and highly undesirable outcome. The presence of a highly undesirable outcome is what will allow us to test the mechanism which previous experimental approximations of the original position have failed to test.

Agenda and Motivation

In the original position, parties are supposed to negotiate the basic terms of their association—principles by which their society should be governed. One could argue that political attitudes which appear under a certain 'association' should not necessarily be interpreted as voices in a negotiation about the terms which should regulate the entire structure of that same 'association'. In other words, the basic terms of the association of the men in the Vietnam lottery are *already settled* and the attitudes that they adopt may be better interpreted as examples of reflections and opinions that can appear in a system which operates under such terms. It is thus questionable whether the expression of their political attitudes can really be understood as the kind of collective negotiation which is supposed to take place behind the veil of ignorance.

The agents behind the veil of ignorance are, however, *supposed* to argue from the point of view of someone whose only known characteristics are those of being self-interested and non-altruistic. Recalling our discussion about Frohlich and Oppenheimer's misinterpretation of the agenda of negotiations in the original position, I argue this to be a strength rather than a weakness. While there is no way of clearly distinguishing how much their political attitudes were defined by their self-interest and their previously held convictions respectively, our material, since it looks at how attitudes change as a consequence of alterations in self-interest, actually provides us with useful tools to make convincing estimates.

Stakes

The vulnerability associated with the threat of ending up as worst off is what can be approximated as the stakes faced by the young college bound men. Receiving a low lottery number presented a very tangible risk of death. Can the conditions of the worst off be approximated to the outcome of potential death? Maybe not. But the 'alternative outcome' for both of them could be formulated as "being able to lead a life in accordance with one's fundamental interests" (Rawls, 1999 :149), and this is, arguably, good enough. With the results of Stoker and Erikson, we can evaluate the extent to which the college bound men's viewpoints changed as a result of this induced vulnerability.

Another way of motivating the similar levels of stakes is to consider the choice between having the life of a young American man in the late 60's, getting drafted to fight in the Vietnam war, or the life of, for example, someone born in the slum of New Delhi. In this hypothetical choice,

we have every reason to seriously doubt that anyone would choose the life in a slum. While the risk of death to many represents the ultimate level of radicalness, and while the Vietnam war was going really bad for the US at the time of the lottery (huge numbers of casualties and political opposition from world-leaders) (Stoker & Erikson 2011:223)— going to war does not necessarily entail death. Instead, many of these men were able to return to the US after the end of the war, and, in many cases, these would therefore still have greater opportunities to 'lead a life in accordance with their fundamental interests' than do many born into poverty. I therefore argue that the stakes of both situations are extremely radical.

However well-matched the potential outcome was of ending up as worst off, my approximation has less to offer regarding the potential sacrifices of ending up as 'better off' (not having to go to war) under the maximin principle. Where I criticised Frohlich and Oppenheimer for not producing a reliable translation of the 'worst off' position to deem it influential enough on participants' choices, the opposite could be said about my own approximation. Soothing the vulnerability faced by the "worst offs" by stopping the war does not seem to come at a cost of those who are 'better off', who are not at risk. They would not, in the case of the war getting stopped, have to sacrifice parts of their 'expected utility' like the case when a maximin principle is adopted in the original position. And while participants' attitudes towards redistribution, as we will see, also changed as a consequence of the lottery draft, the differing natures of the stakes cause problems for the approximation. In the original position, the outcome is described in terms of a solution to a problem of vulnerability, where the solution and the problem belong to the *same category of vulnerability*. The question is: what societal organisation will ensure access to a life lived according to my own convictions, without knowing either my societal position or my convictions? The solution, according to Rawls, is presented in the form of a societal organisational principle which minimises the negative effects associated with ending up as worst off. Since not all of the expressed attitudes in Stoker and Erikson's study, in the same way, coincide with the category of the *cause* of their vulnerability (draft status), we will have difficulties in distinguishing clear variations in Rawls' dependent variable: the principle resulting from negotiations. By examining how subjects' attitudes were affected by the induced vulnerability of the situation, however, I may be able to say something about Rawls' and Hobbes' conjecture; that high enough levels of vulnerability can *generate* a 'universal' and high aversion to risk. As we have seen, this contested assumption has played a crucial role in disagreements about the results of negotiations in the original position.

EVALUATION

Is the approximation of Stoker and Erikson's Vietnam Lottery study to Rawls' original position *good enough* to inform us what principles would be chosen by parties behind the veil of ignorance? While stakes are extremely high, and the outcome uncertain, the subjects' differing levels of uncertainty as well as the different nature of the stakes makes it difficult to reach the

desired level of confidence that we would want to be able to present the results as scientific proofs. And while we, in the results, will discern a clear shift towards preferences for the protection of the 'worst off' in society amongst holders of low lottery numbers, we have no way of clearly distinguishing this as the 'maximisation of income for the worst off'; a 'range constraint'; or a 'floor constraint'. The difficulty in distinguishing clear variations in our dependent variable poses a crucial problem for our approximation of the Vietnam lottery study as an experiment testing *the outcome* of Rawls' original position.

Conclusions about his psychological and behavioural assumptions are, however, interesting enough for the purpose of this paper. The imperfections of our approximation provide no reason to completely discard these results. Below (Table 5), we find a qualitative comparison of the approximation of the Vietnam Lottery study to the experimental approximations of Frohlich and Oppenheimer. Arguably, the differences in quality are not important enough to declare the superiority of one of the approximations over the other. Besides, seeing how their weaknesses and strengths do not coincide, we may even be able to make a joint interpretation of their results to test the feasibility of Rawls' psychological assumptions.

TABLE 5: QUALITATIVE COMPARISON OF EXPERIMENTAL APPROXIMATIONS

Conditions	Rawls	Frohlich & Oppenheimer	Stoker and Erikson
	The original position	**Lab experiments**	**Vietnam Lottery**
Imperfect information	• Natural talents • Conception of the good • Societal position • Economic stage of development of country	• "Income class" i.e. monetary reward • Range of payoffs in 'situation', i.e. the economic stage of development of the country	• Draft status, i.e. determinant of future life chances
Stakes	• Life prospects	• Monetary reward of $6.50-$71.60 • No sufficiently unwanted 'worst off'-position	• Life prospects • No sufficiently attractive 'best off'-position
Agenda and Motivation	<u>Agenda</u> • Settle basic terms of their association <u>Motivation</u> • Self-interested • Non-altruistic	<u>Agenda</u> • Agree on principles of distributive justice for the division and distribution of monetary rewards. <u>Motivation</u> • Self-interested (although clearly influenced by personal convictions)	<u>Agenda</u> • Expression of personal attitudes <u>Motivation</u> • Presumably self-interested

RESULTS[15]

Lottery numbers have been indexed from 1-366 to 0-1. War attitudes as well as different political attitudes have been scaled from 0 (dovish/liberal/Democratic) to 1 (hawkish/conservative/Republican). Since the assignment of lottery numbers was completely random, the difference in attitudes between holders of high versus low numbers is considered to be due to factors associated with holding a low number.

I) War Attitudes

As is visible in Table 6, the students' attitudes towards the war were, to a great extent, determined by the lottery number that they had been given. Between holders of the lowest lottery number and the highest, we can expect a 0.24 increment change in attitudes, meaning that students that were 'safe' from the risk of getting drafted were likely to adopt more positive attitudes to the war. As is also visible, no such difference in attitudes is discernible for men that, prior to the lottery, had not previously been exempt from the draft via a college bound status. Important to remember from Figure 2 (p.25) is that war attitudes never were expected to reach levels as high as 1. As we know, people who were never in danger of getting drafted could still often be opposed to the war. The expected attitudes of holders of the highest lottery numbers, who can be expected not to have felt any risk of getting drafted, seem to have averaged around 0.5. Holders of lottery numbers from around 1-98, on the other hand, who faced the risk of getting drafted, adopted significantly more negative attitudes towards the war, as these range from around 0.22-0.27. Clearly, *actually* being subjected to the risk of getting drafted *personally*, (whether or not they were, in the end!), subjects with lower numbers perceived the situation in a much more negative way than did those who were not.

TABLE 6: EFFECT OF 1969 LOTTERY NUMBERS ON ATTITUDES TOWARD VIETNAM WAR, 1973

(Stoker and Erikson 2011, p.226)

	College Bound	Non–College Bound
	($n = 256$)	($n = 118$)
Lottery number	0.24 (0.07) $p = 0.002$	−0.07 (0.11) $p = 0.550$

Notes: The dependent variable is the composite Vietnam War attitude index, scaled to run from 0 (Dove) to 1 (Hawk). Lottery number is rescaled from 1 to 366 to 0 to 1. Entries are ordinary least squares (OLS) unstandardized coefficients. Robust standard errors (SEs), which take into account the clustering (by school) in the data, are shown in parentheses (see Nichols and Shaffer 2007). Cases are male respondents who had not served in the military as of 1969. "College bound" are those taking college preparatory courses in 1965. *Placebo test results*: Coefficients on lottery number for college-bound women are as follows: −0.00, $p = 0.97$ (bivariate, $n = 295$) and −0.01, $p = 0.88$ (multivariate, $n = 290$).

[15] Unless stated otherwise, all of the below tables will relate to college-bound men only.

II) Political Attitudes

Comparing political opinions expressed by subjects in 1965 with those expressed in 1973, Stoker and Erikson could make estimates of the effects of draft vulnerability on political attitudes. In Table 7, we can see varying levels of correlation between the subjects' attitudes expressed in 1965 with those they expressed in 1973, by how vulnerable they had been to the draft. As we can see, subjects with high numbers are much more likely to stick to the opinions that they had expressed in 1965. Clearly, the exposure to a high risk of getting drafted had the effect of radically changing subjects' political attitudes.

TABLE 7: CORRELATION BETWEEN 1965 PARTY IDENTIFICATION AND 1973 POLITICAL ATTITUDES BY LOTTERY NUMBER: COLLEGE BOUND ONLY

(STOKER AND ERIKSON 2011, P.232)

	Among Those with Low Lottery Numbers (1–122)	Among Those with High Lottery Numbers (245–366)
Correlation of 1965 Party ID with		
Vietnam attitude index	0.07	0.06
1972 Vote choice	−0.03	0.43***
Rating of Nixon vs. McGovern	0.00	0.26*
Partisan political activity	0.11	0.24*
Composite issue index	0.01	0.31**
Political ideology index	−0.05	0.42***
1973 Party ID	0.27*	0.56***

Notes: Correlations were based on pairwise deletion of missing data. Cases are college-bound (those whose 1965 high school curriculum was college preparatory) male respondents who had not served in the military as of 1969. *N*s ranged from 57 to 75 for the low lottery number group and from 60 to 84 for the high lottery number group.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

In Table 8, we can see a summary of how lottery-induced changes in attitudes differed between 1965 identified -Democrats versus -Republicans.[16] Subjects that were holders of the lowest lottery number, and that in 1965 had identified as Republicans, were likely to have 0.31 points more *left*-leaning issue attitudes (on a scale from 0-1) than were previously identified Republicans with the highest lottery number. The corresponding change among previously identified Democratic students was 0.03, which would rather mean for the highest number holders to be slightly *more* left-leaning than previously identified Democrats with the lowest number. This difference is, however, not statistically significant.

Receiving a low lottery number induced right-wing students to adopt left-wing attitudes *in virtually all categories of measurement*, whereas students that had previously identified with the left

[16] All calculations are based on Stoker and Erikson's interaction table (2011:233).

remained leftist. Clearly, low lottery numbers led to the adoption of left-wing political preferences.

TABLE 7: ATTITUDE CHANGE BY LOTTERY NUMBER AND 1965 PARTY IDENTIFICATION

(ALL ENTRIES ARE OLS-REGRESSION COEFFICIENTS)

	1965 identified Democrats, change between lottery number 1-366	1965 identified Republicans, change between lottery number 1-366
Vietnam Attitude Index	0.18	0.31
Rating of Nixon vs. McGovern	0.03	0.31
Partisan Political Activity	0.05	0.21
Composite Issue Attitude Index	-0.03	0.31
Political Ideology Index	-0.10	0.36
Party Identification	-0.14	0.29

APPROXIMATION OF RESULTS TO RAWLS' ORIGINAL POSITION

None of these results were obtained from an experiment which, in any way, was intended to test Rawls' psychological assumptions about agents in the original position. When interpreting them, we must therefore do so with caution. To start off, a *maximal* interpretation of the results is that people, when confronted with the kind of vulnerability which is induced through uncertain and risky conditions of the kind in the original position, adopt preferences which protect them from the worst-case scenario. This would rely on an understanding of left-wing politics—favouring state guarantees of social security and redistribution—as an expression of such preferences. But why would vulnerability towards getting drafted lead to preferences of protective guarantees of material welfare: something totally unrelated?

Stoker and Erikson suggest that the shift towards leftist attitudes could be explained by a combination of factors, including subjects blaming Nixon for the vulnerable position that they were in, since he had implemented the draft system. The shift could also have resulted from subjects' opinions developing to become so against the war itself that the war-related question became decisive for their general political orientation. Moreover, as they participated in anti-war activities, which in turn were organised and attended mostly by the left, we can expect political socialisation to have played a role. One last suggestion is that the vulnerability to the draft induced vulnerability in other aspects as well, as it prevented the 25-26-year old men from making long term commitments, which could have detrimental effects on their life prospects in terms of employment, career prospects, and even family- and love life (2011:230).

We could accept, as a result of more leftist political orientations, an increased preference for protective distributional principles in general. But we still have no way of distinguishing any specific preference for a maximin; floor-; or even range constraining principle. The maximal interpretation subsequently appears only half-convincing and lacking of variation in our dependent variable; the preferred distributional principle.

However, as we compare the ways in which one might argue for the floor constraint- or the maximin principle, we can discern a relevant difference. When arguing for the maximisation of the average with a floor constraint, one attributes importance to *more than one* aspect; one makes trade-offs among the floor, the ceiling, and the mean. Conversely, arguments for the maximin principle only focus on the welfare of those that are worst off.[17] As is visible in the results that Frohlich and Oppenheimer obtained when they raised the stakes of the experiments, these variations had the effect of enhancing the desirability of the principle which took as its *only* concern the maximisation of average. As stakes were perceived as higher, the principle which attempted to cater to more than one consideration, while still the most popular one, lost a significant share of its relative support. Seeing how lab experiments only very marginally can be said to approximate the worst-off position of the original position, it is reasonable to wonder what would happen to the consideration of other factors once this position starts to hit undesirability levels that resemble those imagined by Rawls.

The stakes in the Vietnam Lottery provide a closer approximation to those stakes. Political preferences of subjects with low lottery numbers changed—radically—to fit that of their central concern: not wanting to go to war. This was clearly a consequence of the participants being exposed to the risk of getting drafted. This supports my argument that personal and high

[17] One may want to claim that some 'economic need of incentives for productivity' -types of considerations also are present in the maximin principle, as inequalities clearly are to be tolerated in the case where they are needed for the economy to be productive. But this is only so if that level of productivity is that which *most benefits* those at the bottom. Its value is secondary and purely instrumental; it is not something which Rawls expects people to prioritise, for its own sake, when in the original position.

stakes will have the effect of neutralising the concern that one might, otherwise, attribute to other aspects in negotiations about tradeoffs in the original position, and suggests that restrictions on self-knowledge are unnecessary for the maximin principle to be chosen.

If it is true that the presence of high stakes neutralises all but one consideration—that of self-protection in the worst-case scenario—and if the stakes that are present in the original position are considered to be such stakes, we may want to seriously consider whether the resulting negotiations still can be said to match Frohlich and Oppenheimer's definition of impartial reasoning: "premised on setting aside one's particular interest and perspectives and giving balanced weight to the interests of all" (1992:3). For there to be a match, we must not interpret "balanced weight" as *equal* weight. "Balanced weight" should, instead, be interpreted as 'due' weight, and impartial reasoning would, in effect, entail the determination of what this 'due' should consist in. Rawls suggests that 'due' weight equals attributing *all* the weight to the side of the 'worst offs'. He suspects that when people attempt to reason impartially, themselves, do not put their heads *enough* to the potentially intolerable aspects of ending up as 'worst off', and that they therefore cannot distinguish *just how much* weight ought to be given to that side.

A *minimal* interpretation of the results, hence, refrains from making a *substantial* interpretation of the political attitudes in favor of a focus on the self-protective *mechanism* discerned in the students' sudden change of values. *If* the perceived solution (the end of the war) had been offered by the right instead of the left, we would, in this view, instead have expected to see a massive shift from left-wing to right-wing preferences. The abstention from making a substantial interpretation of the results does not make the minimal interpretation void of substance. Rather, it decidedly points in the direction of Rawls' and Hobbes' assumptions of human nature: when confronted with radical risk, people overlook other, competing, inclinations—all to ensure their future capacity to secure their fundamental interests.

CONCLUSION

In my evaluation of Frohlich and Oppenheimer's experimental approximation, I showed that the resulting preference for the principle with a floor constraint most likely depended on participants, *themselves*, attempting to become 'impartial observers', which hampered the intended functioning of the original position. I argued that their experimental approximations were not good enough, and that this was due to the absence of a sufficiently undesirable worst-case scenario. The undesirability of ending up as worst off in the original position effectively provides the only incentive for people to choose the maximin principle, and in its absence Rawls' assumptions cannot be tested.

I subsequently hypothesised that a sufficiently unwanted worst-case scenario would have been enough to neutralise the influence of people's other convictions which, if true, renders unnecessary the concealing of self-knowledge for the maximin principle to be chosen. To test this hypothesis I made an experimental approximation of my own, using Stoker and Erikson's study on how political attitudes were affected by the Vietnam Lottery. Their results supported my hypothesis that people, when exposed to a high risk, *alter* their convictions to match their central concern: to protect themselves from the worst-case scenario. This, I mean, is independent of individual risk aversion propensities, and should rather be understood as a universal desire to secure our fundamental interests: human nature. This finally leads me to conclude that the maximin principle *is* the likely result of negotiations behind the veil of ignorance, and, thus, for Rawls' empirical assumptions to be valid.

If we agree that the conditions of Rawls' original position—a combination of imperfect information and high stakes—are such as to generate impartiality, my results thus suggest that we also ought to accept his maximin principle to be a legitimate principle of justice.

As a concluding remark I would like to address future researchers who wish to test Rawls' theory experimentally. As I have argued at length, exposure to a *high enough* risk is something which previous experiments have failed to produce. I insist that this is crucial. Low opportunity costs of reward-generating games can simply not be trusted to—even remotely—instil the sense of potential sacrifice which is felt by agents in the original position. Where ethical considerations of scientific research are unlikely to ever allow for acceptable assimilations of such stakes, it may be that approximations could be somewhat ameliorated if opportunity costs were *significantly* increased—to induce *actual* alterational potentials to life prospects.[18] This might allow for the testing of tradeoff propensities between sacrifices of 'worst off' versus 'best off' outcomes. While receiving funding of sufficient size for this may seem unlikely, an experiment is not required to be set in a traditional laboratory setting. In the popular reality TV program *Paradise Hotel,* for example, the distribution of (high!) monetary rewards in their season finale is modeled according to the logics of 'the prisoner's dilemma'. This way, one may be able to solve the funding-problem. Whether or not a televised version of John Rawls' original position would actually provide good enough approximations to test its outcome will, however, be for someone else to evaluate. My hope is, at least, that my Vietnam Lottery -approximation has proven that experiments can benefit from the search for creative solutions.

[18] Frohlich and Oppenheimer have however, to this point, suggested the use of hypnosis (1992:162).

BIBLIOGRAPHY

ARTICLES

Berinsky, Adam. J., & Chatfield, Sara. (2015). 'An Empirical Justification for the Use of Draft Lottery Numbers as a Random Treatment in Political Science Research'. *Political Analysis*, *23*(03), 449–454. http://doi.org/10.1093/pan/mpv015

Chong, Dennis, Citrin, Jack, & Conley, Patricia. (2001). 'When Self-Interest Matters'. *Political Psychology*, *22*(3), 541–570. http://doi.org/10.1111/0162-895x.00253

Erikson, S. Robert., & Stoker, Laura. (2011). 'Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes'. *American Political Science Review*, *105*(02), 221–237. http://doi.org/10.1017/s0003055411000141

Gropp, Reint., Gruend, Christian., & Guettler, Andre. (2014). The Impact of Public Guarantees on Bank Risk-Taking: Evidence from a Natural Experiment. *Review of Finance*, *18*, 457–488. http://doi.org/doi:10.1093/rof/rft014

Mongin, Philippe. (2001). The impartial observer theorem of social ethics. *Economics and Philosophy*, *17*(02). http://doi.org/10.1017/s0266267101000219

BOOKS

Freeman, R. Samuel (2007). *Rawls*. London: Routledge.

Frohlich, Norman., & Oppenheimer, A. Joe. (1992). *Choosing justice: an experimental approach to ethical theory*. London: University of California Press.

Hobbes, Thomas. (1651) 'The Misery of the Natural Condition of Mankind'. In: Rosen, Michael & Wolff, Jonathan (Eds.) (1999), *Political Thought*. chapter, Oxford: Oxford University Press.

Moehler, Michael. (2018). *Minimal morality: a multilevel social contract theory*. Oxford: Oxford university press.

Rawls, John. (1999). *A theory of justice: revised edition*. Cambridge: The Belknap Press of Harvard University Press.

DOCUMENT

Nixon, Richard (1969) *Message from the President of the United States relative to the reform of the selective Service System,* Document No. 91-116, pp. 3-5. 91st Congress 1st Session, House of Representatives. Rettrieved from https://www.rand.org/content/dam/rand/pubs/monographs/MG265/images/webG0671.pdf

PODCAST

Wolff, Jonathan., and Warburton, Nigel. (Producer). (2010, February 28). Philosophy Bites. *Jonathan Wolff on Rawls' A Theory of Justice* [Audio podcast] Retrieved from https://secure-hwcdn.libsyn.com/p/f/3/6/f3677a5c92658ba6/Jonathan_Wolff_on_John_Rawls_A_Theory_of_Justice.mp3?c_id=1779689&cs_id=1779689&expiration=1546410904&hwt=9978d4d8ec506a98b096420f397dc10b

NEWS REPORT

Sutter, John. and Davidson, Lawrence. (2018, December 16). Teen tells climate negotiators they aren't mature enough. *CNN*. Retrieved from https://edition.cnn.com/2018/12/16/world/greta-thunberg-cop24/index.html

APPENDIX

Contents

Frohlich And Oppenheimer's Experimental Design

Educational element

Participants were presented with brief descriptions of the following notions of a "just distribution," after which they were tested on their understanding of the different notions and asked to rank them according to preference (1992, pp. 35-36):

- Maximising of floor: "The most just distribution of income is that which maximises the floor (or lowest) income in the society. This principle considers only the welfare of the worst-off individual in society."
- Maximising the average income: "The most just distribution of income is that which maximises the average income in society. For any society maximising the average income maximises the total income in the society."
- Maximising the average with a floor constraint: "The most just distribution of income is that which maximises the average income only after a certain specified minimum income is guaranteed to everyone."
- Maximising the average with a range constraint: "The most just distribution of income is that which attempts to maximise the average income only after guaranteeing that the difference between the poorest and the richest individuals (i.e. the range of income) in the society is not greater than a specified amount."

PART I

Upon completion of the test, participants were presented with examples of income distributions which could result from the different distributional principles.[1] While there were five income categories (from low-high), participants were unaware of the likelihood of ending up in either of these (1992, p.192). They were given four chits with different 'situations' (ranges of monetary values available in society; their stage of economic development), without explicit categories of distribution principles (these were now referred to as numbers from 1-4).[2] They were asked to indicate their distribution of preference in each situation before they were presented with a new one, and they were told that they would earn $1 per $10,000 of the income of the class that they were randomly assigned with (1992, p.38). Table A1 shows an example of what information was given to participants once they had been assigned to an income class of a situation (in this case, the lowest income class of situation 'A'). As is visible, they were given a summary of the payoffs that they would have gotten if they had chosen any of the other distributional principle (1992, p.39). The keeping of these enabled for participants to look back on previous consequences of their choices and this was assessed, by Frohlich and Oppenheimer, to have had a great pedagogic impact on participants (1992, p.38).

[1] To ensure that due concern is given to the size of the monetary rewards in the experiments, the monetary rewards in this paper are all expressed in terms of 2018 price levels, calculated with a multiplier of 1,79.

[2] To facilitate the reading of these tables, I have chosen to indicate the principle which corresponds to each number. This was, however, not the case in Frohlich and Oppenheimer's experiments.

TABLE A1: INCOME AND PAYOFF PER DISTRIBUTIONAL PRINCIPLE WHEN ENDING UP AS 'WORST OFF' IN SITUATION 'A' (2018 PRICE LEVELS)

Principle of Justice	Income	Experiment payoff
Maximin	23270	$2,3
Maximising the average	10740	$1,1
Floor constraint	18000	$1,8
Range constraint	21480	$2,1

Before they received the chit which presented situation B, they had to decide on their preferred distributional principle in situation A. Participants indicated their distribution of preference in each of the situations, were randomly assigned with a position in that distribution, and their payoff from each of the situations was recorded.

Situations 'A'- 'D'

TABLE A2: SITUATION 'A' 2018 PRICE LEVELS (1992, p.196)

Income Class	1 (Range constraint)	2 (Floor constraint)	3 (Max. average)	4 (Maximin)
High	$50120	$62650	$53700	$44750
Medium high	44750	53700	51910	39380
Medium	35800	44750	50120	34010
Medium Low	26850	26850	48330	28640
Low	21480	17900	10740	23270
Average Income	35800	42065	46540	34010
Floor/Low	21480	17900	10740	23270
Range	28640	44750	42960	21480

TABLE A3: SITUATION 'B' 2018 PRICE LEVELS (1992, p.197)

Income Class	1 (Maximin)	2 (Floor constraint)	3 (Max. average)	4 (Range constraint)
High	$30430	$53700	$71600	$46540
Medium high	28640	44750	53700	42960
Medium	26850	35800	44750	39380
Medium Low	25060	26850	35800	35800
Low	23270	22375	14320	19690
Average Income	26492	34368	42154.5	37411
Floor/Low	23270	22375	14320	19690
Range	7160	31325	57280	26850

TABLE A4: SITUATION 'C' 2018 PRICE LEVELS (1992, p.198)

Income Class	1 (Maximin)	2 (Max. average)	3 (Floor constraint)	4 (Range constraint)
High	$179000	$62650	$53700	$42960
Medium high	53700	53700	44750	41170
Medium	35800	44750	41170	39380
Medium Low	26850	35800	26850	37590
Low	23270	14320	21480	19690
Average Income	34905	40006.5	35800	37070.9
Floor/Low	23270	14320	21480	19690
Range	155730	48330	32220	23270

TABLE A5: SITUATION 'D' 2018 PRICE LEVELS (1992, p.199)

Income Class	1 (Floor constraint)	2 (Max. average)	3 (Range constraint)	4 (Maximin)
High	$62650	$53700	$35800	$53700
Medium high	53700	50120	32220	50120
Medium	44750	46540	28640	42960
Medium Low	35800	42960	25060	35800
Low	23270	21480	21480	25060
Average Income	42154.5	43855	27745	40633
Floor/Low	23270	21480	21480	25060
Range	39380	32220	14320	28640

PART II

In part II, participants entered discussions about the distributional systems with the other participants until they, unanimously, could decide on a distributional principle of preference. They were told that if they failed, both the distributional principle and their social position, which combined determined the size of their payoff, would be assigned at random (not only their own assigned income class, as would be the case if they came to an agreement). The ideal was, of course, for the distributional principle to be chosen by the participants and for only their social position to be chosen at random. Unanimity was ensured by a double voting procedure where the first one was open and the latter anonymous through the casting of ballots. Participants were told that the stakes in Part II were significantly higher, but they were not told how these were to be calculated (1992, p.200). When they had decided on a distributional principle, they were randomly assigned with an income class with a corresponding monetary payoff that belonged to a 'situation' which was also selected at random.[3]

In table A6, the highest and the lowest incomes, as well as the highest and the lowest payoffs in each of the situations (A,B,C, and D) are summarised (1992, pp.196-199).

[3] Some experiments later went on to test the stability of the chosen principles, but this falls outside the scope of this paper.

TABLE A6: ANNUAL INCOMES OF LOWEST VS HIGHEST INCOME GROUPS IN THE SITUATIONS PRESENTED TO PARTICIPANTS (PAYOFF WITHIN BRACKETS) 2018 PRICE LEVELS (*HIGHEST/LOWEST)

		Maximisation of average income	Maximisation with floor constraint	Maximisation with range constraint	Maximin
Situation A:	Lowest	$10,740 ($1,1)*	$17,900 ($1,8)	$21,480 ($2,1)	$23,270 ($2,3)
	Highest	$53,700 ($5,4)	62,650 ($6,3)*	$50,120 ($5,1)	$44,750 ($4,5)
Situation B:	Lowest	14,320 ($1,4)*	22,375 ($2,25)	19,690 ($2)	23,270 ($2,3)
	Highest	71,600 ($7,6)*	53,700 ($5,4)	46,540 ($4,7)	30,430 ($3,4)
Situation C:	Lowest	14,320 ($1,4)*	21,480 ($2,1)	19,690 ($2)	23,270 ($2,3)
	Highest	62,650 ($6,3)	53,700 ($5,4)	42,960 ($4,3)	179,000 ($17,9)*
Situation D:	Lowest	21,480 ($2,1)*	23,270 ($2,3)	21,480 ($2,1)*	25,060 ($2,5)
	Highest	53,700 ($5,4)	62,650 ($6,3)*	35,800 ($3,6)	53,700 ($5,4)

Motivational Calculations of Estimates

'Best Off'

The biggest reward that one could leave the experiment with was $71.60 (Frohlich and Oppenheimer 1992, p.45).[4] In order to receive the maximum reward, participants would repeatedly have to choose the distributional principle which maximised the income for the highest income class, and subsequently be fortunate enough to end up in that class. If one managed with this in all four situations presented to the subjects in Part I (A, B, C,

Table A7: Effect of ending up as 'best off', principle by principle (payoff within brackets, highest payoff in bold)

	Maximisation of average income	Maximisation with floor constraint	Maximisation with range constraint	Maximin
Situation A	$53,700 ($5,4)	**62,650 ($6,3)**	$50,120 ($5,1)	$44,750 ($4,5)
Situation B	**71,600 ($7,6)**	53,700 ($5,4)	46,540 ($4,7)	30,430 ($3,4)
Situation C	62,650 ($6,3)	53,700 ($5,4)	42,960 ($4,3)	**179,000 ($17,9)**
Situation D	53,700 ($5,4)	**62,650 ($6,3)**	35,800 ($3,6)	53,700 ($5,4)
Total (Part I)	$24,70	$23,40	$17,70	$31,20
Total (Part I+II)	$62,80	$61,50	$55,80	$69,30

[4] $40 in 1992 price levels

and D), one would have accumulated a total of $38.10 at the moment of completion of Part I (Table A7). This means that the maximum reward from Part II must have been $71.60 - $38.10 = $33.50.

'Worst Off'

They do not mention what the smallest possible reward was, but using the figures that they have provided us with, we should be able to make an estimate. If a participant repeatedly chose the distributional principle where the lowest income class had the lowest income[5] (compared to the other distributional principles in that 'situation'), and had the misfortune of ending up in the lowest income class in all four situations, he or she would have accumulated only $6 at the end of Part I (Table A8). If group discussions resulted in a choice of principle which, again, led to an insignificant reward for those in the lowest income class, and if the participant, again, was unfortunate enough to end up in that class, he or she would, again, be given only a small addition to their payoff. While we do not know how small Frohlich and Oppenheimer made the smallest reward, we may expect it to be smaller than the smallest ones in Part I. After all, they specifically said that "stakes", in part II, "were much higher", which should entail that both gains and 'losses' (opportunity costs) were of greater importance (1992, p.200). For the sake of making an estimate, we can suppose that the smallest possible reward that one could get from the second part of the experiment was $0.50, independently of the distributional principle.[6] This would mean that the minimum reward for participating in the experiment, if one were continuously to end up as 'worst off ' under the application of the principle with the lowest income for the lowest income class, was $6.50.

Table A8: Effect of ending up as 'worst off', principle by principle (payoff within brackets, smallest payoff in bold) 2018 price levels

	Maximisation of average income	Maximisation with floor constraint	Maximisation with range constraint	Maximin
Situation A	**$10,740 ($1,1)**	$17,900 ($1,8)	$21,480 ($2,1)	$23,270 ($2,3)
Situation B	**14,320 ($1,4)**	22,375 ($2,25)	19,690 ($2)	23,270 ($2,3)
Situation C	**14,320 ($1,4)**	21,480 ($2,1)	19,690 ($2)	23,270 ($2,3)
Situation D	**21,480 ($2,1)**	23,270 ($2,3)	**21,480 ($2,1)**	25,060 ($2,5)
Total (Part I)	$6	$8,45	$8,2	$9,4
Total (Part I+II)	$6,5	$8,95	$8,7	$9,9

'Average Income'

The calculation of total payoffs in the case where a participant, repeatedly, ended up with the average income (per distributional principle and situation), requires a higher degree of speculation. The relative size of the average income to the highest income varied significantly between both distributional principles and situations (Table A9). To make

[5] These were, as visible in Table A8, all found under the principle which maximised the average income.

[6] It may have been a better assumption that the smallest reward differed between the principles, but this would also require a higher degree of speculation.

things simple, I have therefore assumed the relative size of average income to the max. income to have been *the same for part II* as it was *for the totals of Part I*. Because of some outliers this may, to some extent, be said to have distorted the estimates. For example, the average income, under the principle which maximised average income for situation 'C' (22%), is less than half as big as the second to smallest one of that column. This results in average income, for that principle, being estimated to stand for only 45,1% of max income although it, in most situations, had a greater relative size to the highest income. Outliers of this kind are, however, present under all principles. This, at least, means that we should not expect the distorting of estimates to affect any of the principles disproportionately.

Table A9: Effect of ending up with 'average income', principle by principle, (payoff within brackets, percentage of max income of situation in red) 2018 price levels

	Maximisation of average income	Maximisation with floor constraint	Maximisation with range constraint	Maximin
Situation A	$46,540 ($4.7)74%	$42,065 ($4.2) 67%	$35,800 ($3.6) 57%	$34,010 ($3.4) 54%
Situation B	42,155 ($4.2) 55%	34,368 ($3.4) 44%	37,411 ($3.7) 49%	26,492 ($2.6) 34%
Situation C	40,007 ($4) 22%	35,800 ($3.6) 20%	37,071 ($3.7) 21%	34,905 ($3.5) 20%
Situation D	43,123 ($4.3) 68%	42,155 ($4.2) 67%	27,745 ($2.8) 44%	40,633 ($4.1) 65%
Total Payoff part I (when max $38,1)	$17.2 45,1%	$15.4 40,4%	$13.8 36,2%	$13.6 35,7%
Total Payoff Part II (when max $33,50)	$15.1 45,1%	$13.5 40,4%	$12.1 36,2%	$12 35,7%
Total Part I + Part II (when max $71.6)	$32.3 45,1%	$28,9 40,4%	$25.9 36,2%	$25.6 35,7%