STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's thesis in economics
Academic year 2019-2020

Strong reciprocal judgments: Differences in approval for welfare provision based on recipient deservingness cues

Markus Jury (41216)

Abstract. This thesis examines whether simple deservingness cues can shift individual opinion on the provision of welfare to a hypothetical welfare recipient. Using a pre-registered design and analysis plan based on earlier work by Lene Aarøe & Michael Bang Petersen, I conduct an experiment using students recruited by email from the Stockholm School of Economics with one control and two treatment groups. The control group received no cues about the hypothetical welfare recipient. The two treatments were given either a cue signalling deservingness, termed an 'unlucky cue', or a cue signalling undeservingness, termed a 'lazy cue'. Results were obtained first across all three treatments and showed that there was a significant difference in responses between the three experimental groups. Pairwise comparisons showed differences in responses between the 'lazy cue' group and the other two but could not reject the null hypothesis of no difference in responses between the 'no cue' and 'unlucky cue' groups. I assert that these results have implications for how to conduct experiments on deservingness cues and may imply that subject predispositions influence the outcomes of such experiments.

Keywords: Deservingness, Strong Reciprocity, Social Welfare, Redistribution, Public Opinion

JEL: D63, D83, D91, H50, Z13

Supervisor:	Magnus Johannesson
Date submitted:	11 May 2020
Date examined:	27 May 2020
Discussant:	Linnea Englund Davidsson
Examiner:	Kelly Ragan

Acknowledgements

I would like to extend my heartfelt thanks to the following people: Magnus Johannesson, for his guidance, patience and advice during the thesis-writing process. Anna Dreber Almenberg, for her role as an understanding and especially helpful program director. Lastly, my family, friends and partner for their unwavering support throughout my studies.

Table of Contents

1. Introduction

Social welfare programs are a fact of modern life. Public social spending across the OECD averages a massive 26% of GDP (Barr, 2020). The extent of such programs and more particularly for this paper the support for their implementation and continuation varies significantly between and within populations across the globe, however. Numerous explanations have been forwarded across the fields of economics, psychology, and political science as to why this discrepancy exists. As with many complicated policy matters, no one theory can explain the entire matter, but each contributes to a better understanding of the true mechanisms at work.

Why does public support for such programs matter? Discussing government policy in the case of nudges, Sunstein (2016) notes "no public official will entirely disregard a strongly felt moral concern on the part of significant segments of the public" (p. 182) when deciding on policy implementation. Indeed, public support is often crucial for creating the base of political support needed to enact and continue such programs, especially in the face of sustained and exaggerated attacks by politicians denigrating them (S.A. Levitan, 1985). This threat to public support is exacerbated by a trend of 'destructuration' in modern times, undermining society-wide collectivity and solidarity that often underpinned the support for welfare programs; this drop in support consequently lead to the decline of strong welfare states throughout the western world (Offe, 1987).

Regarding differing support for welfare in particular, neoclassical economics contains two main schools of explanation. In public choice theory, coerced and voluntary redistribution suggest collective pressures for redistribution are what drive its occurrence, and that the motivation behind these pressures are attributable to either the self-interest of those desiring transfers (in the coerced case) or of those desiring optimal utility outcomes (in the voluntary case) (Barr, 1998). Social insurance theory conceives of redistribution as something individuals support because they treat it as an insurance against risks they personally face. While there are reasons for supporting public rather than private provision of such insurance, this support ultimately occurs because such public provision can be more efficient, and is not due to factors beyond individuals acting in a self-interested manner. More recent work in the social insurance field such as Moene & Wallerstein (2003) somewhat expanded the conception of social insurance beyond a mere risk hedge, but retain the focus on self-interested rationality explaining support (or lack thereof) for welfare programs.

Behavioural economics as a whole does not dispense with the self-interest framework entirely, but does move beyond it in important ways. Theories of altruism propose charitable giving or support for redistribution are a result of a complex interplay of self-interested 'warm-glow' benefits from giving (Andreoni, 1989), dispositional factors and social norms (Folbre & Goodin, 2004). Such norms and social expectations can also drive support for redistribution by acting upon opinions concerning redistribution specifically, supported by the evolution of those norms as discussed by Ostrom (2000), Young (2015) and, specifically in the case of welfare opinions, Barón, Cobb-Clark and Erkal (2015). The motivated beliefs framework summarized by Benabou & Tirole (2016) offers a key explanation for why individuals' opinions regarding support for welfare provision may be influenced by cues. Relatedly, repeated messaging that individuals engage with on a low-attention basis can have a strong effect on belief formation (Hawkins, Hoch & Meyers-Levy, 2008). Political cues about policy can strongly affect individuals' belief formation, to the point of increased opinion-formation effort (Petersen et al., 2013). Finally, Krueger & Rothbart

(1988) pioneered a critical branch of literature showing that cues about a person's behaviours can have a significant impact on judgements about that person by others.

Political science and sociology have also offered up a large variety of explanations for gaps in welfare support. With one foot still in the realm of behavioural economics, Alesina, Glaser & Sacerdote (2001) note that the concept of strong reciprocity is fundamental to much of the discussion of support for welfare in these disciplines. This concept proposes that people will go out of their way to help others that display strong reciprocity, while punishing those who are viewed as breaking this norm (Fong, Bowles & Gintis, 2006). This important component in the economic comparative welfare literature is mirrored by the political science and sociological concept of 'deservingness' (van Oorschot, 2000). Ethnic heterogeneity also seems to play a major role in the determination of support for welfare in societies where it is present (Freeman, 1986; Gilens, 1995; Alesina, Glaser & Sacerdote, 2001).

A final crucial part of the literature in the realm of sociology, political science and economy is that of institutional factors. The cornerstone of comparative welfare literature considering institutional path dependencies is Esping-Andersen's (1990) categorization of the three types of welfare state: liberal, conservative, and social-democratic. While his work has been strongly challenged since its conception due to a lack of empirical backing (Papadakis & Bean, 1993), more recent researchers in that framework have used more multidimensional theoretical (Larsen, 2008) and empirical (Jaeger, 2009) approaches with some success. This type of approach is critical when individuals' evaluation of welfare states is often multidimensional (Roosma, Gelissen & van Oorschot, 2012). Other institutional factors may also play a role in determining public support for welfare states, such as the breakdown of collective structures in modern societies (Offe, 1987), constitutionally defined welfare rights (Iida & Matsubayashi, 2010), a cleavage between religious and irreligious citizens (Stegmueller et al., 2012), and macroeconomic factors such as economic crises (Sihvo & Uusitalo, 1995; Kam & Nam, 2007; Margalit, 2013; Soroka & Wlezien, 2014)—though the verdict on this last factor is decidedly mixed.

This paper investigates a particular explanation which draws on both the behavioural/psychological and sociological/political strains of welfare literature. Aarøe & Petersen (2014) forwarded the hypothesis that simple cues about the deservingness of welfare recipients can create shifts in opinion about welfare provision against individuals' natural dispositions on the matter, to the point of crowding out national stereotypes. They conducted a cross-national study with large online survey panels, and found that the introduction of simple cues into a Likert-scale question on welfare support did in fact drastically shift opinion on welfare provision, crowding out national stereotypes as expected.

This thesis seeks to follow-up on their cross-cultural study by investigating if the results on deservingness cues they show can be replicated in a single population: in this case, students at the Stockholm School of Economics. Replication is a key concept to further knowledge in social sciences, and an important hedge against a prevalence of results with low statistical power and issues with so-called "researcher degrees of freedom" (Dreber & Johanneson, 2019). While budgetary and time constraints precluded a proper replication of Aarøe & Petersen's methods, particularly the cross-national and stereotype portions, I intend for this contribution to nevertheless help support or question their general results regarding the effects of deservingness cues on support for welfare provision.

Experimentally, I sought to answer the following question: can the introduction of simple deservingness cues about a hypothetical welfare recipient spur differences in responses to a question about whether social welfare benefits should continue to go to recipients like the hypothetical person?

The hypotheses tested in this paper were as follows: primarily, whether a systematic difference in approval for welfare provision exists depending on deservingness cues presented to respondents. Secondarily, whether a systematic difference in approval for welfare provision exists between specific cues (or lack thereof) presented to respondents. I expected to reject the null hypothesis of no systematic difference in all cases.

This study was conducted online, via the platform Qualtrics. I recruited from a pool of 1700 students at the Stockholm School of Economics via email, and managed to attain a sample size of 320. The participants were asked to complete a short survey, consisting of one primary experimental question following a brief consent request and email confirmation page. They were randomly assigned by the Qualtrics software to one of three experimental groups by being shown only the relevant experimental question. The first group was a control with no deservingness cue (the 'no cue' condition). The other two groups were treatment groups—one with a cue signalling deservingness, the other with a cue signalling undeservingness (the 'unlucky cue' and 'lazy cue' conditions).

In order to strengthen the credibility of my work and prevent issues such as p-hacking, I pre-registered my analysis plan on the Open Science Framework (OSF)[1]. Pre-registration offers increased validity in research, particularly for replication efforts (Dreber & Johanneson, 2019). While my pre-registration did contain some minor errors regarding interpretation of results and I changed the title of my paper since, throughout the experiment I followed my design and analysis plan as listed in the pre-registration. The hypotheses listed in the plan mirror those laid out in this paper, along with the expectation of reject the null hypotheses in all cases.

My results rejected three out of the four proposed null hypotheses, and failed to reject one. I was able to reject the overall null hypothesis that no systematic difference in approval for welfare provision exists depending on deservingness cues. I was also able to reject the null hypotheses that the 'lazy cue' condition did not differ from the 'unlucky cue' and 'no cue' conditions. I was unable to reject the null that the 'unlucky cue' condition differed from the 'no cue' condition.

This paper begins with a Literature Review (Section 2), seeking to provide a general overview of the literature regarding the determinants of support for welfare programs and redistribution, as well as the literature regarding deservingness cues more specifically. It covers the broad areas of neoclassical economics, behaviour economics & psychology, and political science, economy and sociology. Section 3 follows with an explanation of my experimental design, from treatments through subjects, experimental procedure, measured variables, and statistical testing. Section 4 discusses what results I expected to find, as well as listing experimental limitations I faced. Section 5 proceeds to present and analyse the results of my experiment. Section 6 then discusses and evaluates the implications of these results. Section 7 concludes with a summation and a brief discussion of further research opportunities.

[1] A full copy my pre-registration is attached in the appendix. It and my experimental data and code are also accessible at https://osf.io/cftuj/

2. Literature Review

As this thesis is primarily a replication of Aarøe & Petersen's findings regarding the effect of deservingness cues on support for welfare, this review does not delve deeply into any one explanation of why support for welfare occurs. I instead opt to give a curated overview of what various strains of academic thought have to say on the matter to support discussion of how and why such cues may influence the support for welfare programs.

I divide my discussion of the literature surrounding the topic into three sections: the first broadly dealing with the neoclassical economic explanations for this support, the second with behavioural and psychological explanations, and the last with political science, political economy and sociology. Overlap between these sections is of course inevitable—for example, neoclassical economists such as Arrow (1963) discuss how institutions play a part in social insurance formation, and political scientists such as Kam & Nam (2008) investigate how macroeconomic conditions factor into public opinion.

2.1 Neoclassical Economics

Neoclassical Economics engages with the question of why individuals support welfare programs, and redistribution more generally, with a number of approaches that are united in their focus on rational individual action. Public choice theory focuses on demand for redistribution by voting coalitions that would benefit from such redistribution. Traditional rational choice models assert that approval or disapproval of welfare programs is rooted in calculated self-interest, often in the form of considering them as a type of social insurance.

2.1.1 Coerced & Voluntary Redistribution

A primary theory as to why redistribution occurs in the first place (and thus the motivations behind it) in the neoclassical tradition is that of coercive redistribution, as pioneered by Anthony Downs in his treatise "An Economic Theory of Democracy" (Tullock, 1971). According to Downs, politicians seek votes, thus they choose policies that maximize the number of votes they get at the next election; as the income distribution in most societies has the largest (nominal) voting base among the poor, the poor will 'force' redistribution from the rich to them (their highest-expected-utility action) by voting for politicians that pursue such policies (Barr, 1998).

Tullock expanded on this theory by noting that things are rarely that simple. In Tullock's view, redistribution primarily occurs because interest groups pressure politicians into redistributing to them (1970). These groups consist not only of only the poor, but also special interest voter coalitions often featuring a large portion of the middle class; the simple "desire to receive transfers" is the primary motivator behind demand for redistribution, not high-minded charity or envy (Holcombe, 2018). Tullock further noted that these interest groups may well siphon redistribution away from its nominally intended targets through median voting or regulatory capture (Barr, 1998). This suggests that rent-seeking behaviors may distort opinion on redistribution away from what it would be if it achieved its nominally stated aims. It is no coincidence that Tullock's contemporary work on pioneering the modern conception of rent-seeking drew upon the same strand of public choice theory (1967).

Barr (1998) contrasted these approaches with the idealistic views of Hochman & Rodgers (1969): that even individuals who would be providing net transfers might engage in redistribution voluntarily, due to its society-wide pareto optimality based on patterns of interdependent utility. This argument that "progressivity… may be interpreted as a matter of revealed preference" (Hochman & Rodgers, 1969, p.556), while heartening, is undercut by the possible existence of free-riders among the rich, as well as the extent to which the rich prefer redistribution not adding up to the total redistribution seen as desirable or optimal in societies (Barr, 1998).

These income-class theories of the causes of redistribution, while ultimately still devolving to individual self-interest, retain a collective perspective missing from the more libertarian bent of explanations provided by social insurance theories.

2.1.2 Social Insurance

The social insurance literature often, but not exclusively approaches the question of the motivations behind welfare policies on the assumption that individuals' evaluation of welfare programs is based entirely or nigh-entirely on calculated self-interest: considering welfare programs as a form of 'social insurance' for the individual.

An question that immediately follows is: if an insurance market is what redistribution efforts amount to, why do they exist as public redistributive efforts rather than private insurance markets? Barr (1998) brought up three primary reasons why inefficiency and possible lack of supply for certain kinds of insurance means they can nevertheless be served by public redistribution in this framework.

First, informational asymmetry in the form of adverse selection—Akerlof's (1970) 'lemons'—can make insurance markets for things like medical insurance function inefficiently due to consumer behaviour, as buyers know more about their own health than insurers do and can thus take policies that would not be sold to them otherwise. A further issue regarding informational asymmetry is that of moral hazard: insurance-purchasing individuals may under-invest in preventative activities, even if doing so would significantly lower costs in aggregate (Barr, 1998).

Secondly and relatedly, informational asymmetry may also harm consumers to the benefit of insurers: when consumers are not well-informed about the benefits, details and costs of insurance—especially when complex contracts are involved—can be prohibitively high, and reduce or eliminate the benefits of private competition (Barr, 1998).

Finally, high administrative costs—marketing, processing, reimbursement, or foregone economies of scale—may render private insurance coverage inefficient if costs outweigh welfare gains from these excess costs or simply if social provision would be much cheaper (Barr, 1998).

Each of these are an argument for social insurance. Arrow (1963) contended that the numerous information and market failures seen in the medical care industry—such as moral hazard, numerous payment forms, third-party payment control, administrative costs, cost unpredictability, unequal risk pooling and coverage gaps—necessitated other institutions to compensate for the failure of markets, which then sprang up in response to these failures. Institutional factors and the lack of efficient market supply are what underpin the appearance of social insurance in these explanations.

In contrast to Arrow's institutional approach, individual-focused neoclassical views of social insurance argue that support for redistribution grows as median income falls, as seen in Meltzer & Richard (1981). Discussing a "rational theory of the size of government", Meltzer & Richard asserted that voters are well-informed about the costs of social insurance, and opt for it due to a clear-eyed view of the mean income relative to their own (1981). Using a general equilibrium model, they showed that a decisive voting individual (in a fully democratic context, the median voter) will increase redistribution if the mean income in society is above theirs, and lower it if it is below (Meltzer & Richard, 1981). In other words, as an income distribution skews negatively, redistribution through the ballot box ('an increase in the size of government') will occur as the decisive median voter will see it is in their best interest and act upon it.

While Meltzer & Richard's model relied on the strong assumptions that voters are informed and empowered enough to actively 'choose' taxation levels, extensions such as Moene & Wallerstein's 2003 incorporation of a social insurance explanation also follow the self-interest route laid down by Meltzer & Richard without demanding such strong assumptions.

Moene & Wallerstein built on earlier social insurance literature and the redistributive arguments made by Tullock and others by rejecting a binary between risk insurance and a self-interested desire for redistribution, arguing that both motivations are what spur social-insurance policies into existence (2003). They further asserted that this mixture of motivations is what determines the relationship between inequality and welfare support, and that this helped explain why their empirical work showed some welfare spending actually declining as income inequality increased (Moene & Wallerstein, 2003). In essence, Moene & Wallerstein's model predicted that an increase in inequality will lower the demand for redistribution in the form of income insurance, as the median voter's income has fallen and thus there is less need to insure against the risk of losing that income (2003). However, it would also increase the demand for redistribution more generally to compensate for the gap between the median and mean income as per Meltzer & Richard (1981). Empirically, Moene & Wallerstein asserted that they found only categories of welfare spending where rising inequality reduced aggregate demand for redistribution (e.g. unemployment benefits) and where they had no effect (e.g. pensions, health care) (2003). They found no categories where rising inequality increased demand, because by their model such a policy would have to redistribute purely to active participants in the labor market (Moene & Wallerstein, 2003).

While these neoclassical economic theories as to why redistribution occurs differ in their focus on collective versus individual power, they all refer back to rational self-interest as the primary motivator for support for redistribution. Going by their arguments, the perceived characteristics of other recipients of welfare should make little to no difference to individuals' support of such programs. With the exclusion of Hochman & Rodgers' (1969) interdependent utilities, neoclassical investigations generally do not consider individuals' perceptions of welfare recipients in their calculations. Even Hochman & Rodgers only approach social relationships with an arms-length interdependent utility framework. As Alesina, Glaser & Sacerdote noted in their comprehensive comparative welfare state survey, rational-self-interest explanations "do not explain much of the puzzle" of the differences in redistribution or support for it between Europe and the United States (2001, p.246). Empirical work by researchers like Papadakis & Bean (1993) agreed that self-interest was 'overemphasized' as a determinant of shaping opinion about public services. To investigate these factors further, one must turn to the literature in behavioural economics and psychology.

2.2 Behavioural Economics & Psychology

While behavioural economics arguably asserted itself enough in the modern day to be counted as part of 'orthodox' economics, I separate it into this section due to the its shift in focus when it comes to explaining public support for welfare. Rational self-interest is still a focus of some behavioural models (particularly that of warm-glow altruism), but many go beyond it, drawing on psychological literature. Altruism is, however, one of the most common explanations for support for redistribution, particularly for support from those who do not directly benefit from the provision of welfare.

2.2.1 Altruism

The idea that altruism is a motivation behind support for redistribution originates long before more recent behavioural advances in experimental economics. Tullock noted in 1970 that "the common view among modern intellectuals [is] income redistribution is considered to be a rather simple and almost entirely ethical matter… those of us who are well-off use the state as a mechanism for making gifts to the poor" (p.379). In his later work, he would continue to assert that while it constituted a secondary motive, charity nevertheless was subordinate to the effect of the desire of special interest groups to receive transfers in accounting for most redistribution that occurs (Holcombe, 2018). Be that as it may, Tullock's grudging admission that altruism did play a role in supporting redistribution indicates it is worth bringing into the general discussion.

More recent economic literature has greatly expanded the detail with which altruism is examined in economics, with James Andreoni's work on the concepts of pure and impure altruism (and their measurement) at the forefront of this wave. Andreoni's conception of impure altruism placed much importance on the existence of a 'warm-glow' effect, in which charitable donors receive more utility from the act of giving away wealth than they would holding on to this wealth themselves (Andreoni, 1989). Andreoni's work, however, limited itself largely on the effects of altruism on private philanthropy and the effect of taxation policy on such philanthropic spending. Andreoni & Miller's experimental investigation of altruism as a matter of revealed preferences is relevant to this thesis both for its discussion of the concept of 'crowding out', as well as the assertion that individuals' altruism is heterogenous, with what they assert are preferences for fairness and inequality-aversion ranging widely (2002).

This focus on rational charitable giving constitutes a significant portion of the economic literature on altruism, but certainly not its entirety. It is instructive in this instance to refer to a discussion of altruism in the political scientist K.R. Monroe's 1996 book *The Heart of Altruism*. Discussing 'Economic Approaches to Altruism', Monroe concluded that traditional preference approaches like Andreoni's have pertinence to rational volunteer or charitable giving activities, and some regarding philanthropic activities, but fail to explain more extreme, especially self-sacrificing forms of altruism. According to Monroe, different, less individualistic conceptions of altruism are required to explain the true breadth of human behaviour in this area.

Though still approaching the topic in an overarching rational actor framework, Folbre & Goodin discussed issues that crop up when altruism is conceived as a matter of individual revealed preferences (2004), in the spirit suggested by Monroe. Their primary focus was on the dilemma of mutual altruism, where the assumption of the inscrutability of the mind to others comes into conflict with the theoretical necessity of each altruist to incorporate the others' preferences into

their own preference function (2004). They concluded that the motivations behind altruism are a complex mix of factors that are best captured by treating altruism not just as a matter of revealed preferences, but rather a dynamic matter of 'dispositions' (Folbre & Goodin, 2004). Dispositions are not fixed, settled, or easily summarized by a revealed-preference framework, and manage to incorporate important factors such as the potentiality of altruistic behaviour, reciprocity, and norms (Folbre & Goodin, 2004). This conception of dispositions was further explored by Guala (2019), who argues preferences themselves are constrained by prior definitions and are better conceived and discussed as "belief-dependent dispositions" (p. 396).

The actual link between altruism as a motivator for the support for welfare programs is generally treated as straightforward: providing support for redistribution in the form of welfare programs is considered an altruistic act on the part of those who would pay more for such programs than they would gain from them. While there are certainly more factors at play here, as Folbre & Goodin's (2004) discussion of complex altruism captured, this generalized relationship is a sufficient reason to include altruism in the discussion of public support for welfare programs.

Another important influence on the support for such programs that Folbre & Goodin (2004) discussed briefly is that of norms and social expectation, noting that ultimately a strong enough normative order can blend altruistic and egoistic motivations to result in the same altruistic behaviour.

2.2.2 Norms & Social Expectation

The issue of normative influences on support for welfare programs can be approached indirectly, in terms of norms of altruism, but it is not difficult to conceive of a more direct normative order where a norm of expected levels of redistribution drives that support. Norms definitionally regulate behaviour, and such a norm or set of norms could help explain some of why individuals support welfare programs. Ostrom (2000) and later Young (2015) both discussed at length factors that contribute the development and endurance of social norms, from addressing collective action issues, the effects of chance events, norm persistence or the compression of behaviour effected by norms.

Barón, Cobb-Clark and Erkal examined in more detail a route for the evolution of norms supporting welfare programs (or not) through the possibility of intergenerational transfer of such norms (2015). Their work showed a significant effect of socialization on young people's approval of generous social benefits: growing up in a family with a history of welfare receipt increases the likelihood of approval, while having mothers that disapprove of welfare or were employed reduces that likelihood (Barón, Cobb-Clark and Erkal, 2015).

Social expectation and judgements bridge the gap between norms and beliefs, and Krueger & Rothbart (1988) attempted to show experimentally that cues about a person's supposed traits can have a significant impact on judgements about that person by others. Particularly relevant for the type of cues used by Aarøe & Petersen (2014) and my own work, Krueger & Rothbart's (1998) second experiment suggested that when behaviours were depicted as temporally consistent rather than one-off, such descriptions dominated other influences on resulting trait judgements.

2.2.3 Motivated Beliefs

Benabou and Tirole (2016) examined another behavioural concept relevant to this topic, that of motivated cognition, and in particular motivated beliefs. Their work discussed how motivated beliefs can lead to situations where, when objective information is scarce, identity-enhancing behaviours are "easily affected by minor manipulations of salience such as cues, reminders, and semi-transparent excuses" as individuals adjust their behaviours and beliefs to fit their self-image (Benabou and Tirole, 2016, p. 159). This could, for example, entail condemning welfare programs due to seeking to reinforce ones' own identity as a hard-working citizen who does not rely on them in the face of cues about free-riding welfare recipients. Benabou & Tirole (2016) went on to discuss how these beliefs can lead to ostracism and condemnation of those seen to be violating norms. A possible example applicable to the topic of this paper would be ostracism of perceived free-riders benefiting from redistribution programs.

Following in the path laid down by the work of Herbert Krugman, Hawkins, Hoch & Meyers-Levy (2008) showed in the context of product advertising that repeated exposures to messaging, particularly when it is not scrutinized carefully, can lead to large impacts on consumers' beliefs about products. They pointed out that the deluge of messaging in the modern era has only intensified this effect, as people are exposed to tens or hundreds of low-involvement messages a day (Hawkins, Hoch & Meyers-Levy, 2008). Petersen et al. (2013) showed that partisan cues on political positions led participants to engage in opinion formation of motivated beliefs; in other words, messaging or cues about policies (such as welfare programs) by politicians or political parties can substantially influence citizens' opinions. Indeed, the "mere presence of party cues triggers increased effort when an opinion is formed" (Petersen et al., 2013, p. 849), meaning that citizens may change their opinions in response to cues despite countervailing inherent beliefs or predispositions.

The topic of motivated beliefs has particular salience for the next section of the literature review: that of political science, political economy & sociology. These fields all attempt to study how humans interrelate, often in the realm of beliefs and norms, and in so doing they undoubtedly fall prey to many of the motivated reasoning errors detailed by Benabou and Tirole.

2.3 Political Science, Political Economy & Sociology

Political and sociological explanations for public support for redistribution take the shift in focus from individual to collective motivation even further than behavioural economics. From the behavioural-economics-adjacent concept of strong reciprocal altruism through the effects of racial discrimination, institutional and cultural influences, and macroeconomic impacts on societies this branch of the literature largely deals with causes outside of the control of any one person.

2.3.1 Strong Reciprocal Altruism

Bridging disciplines between behavioural economics and political science, the concept of strong reciprocal altruism or 'deservingness' is another key variable for explaining public support for welfare programs. As it forms the backbone of Aarøe & Petersen's experiment, it is also particularly crucial for this thesis.

Fong, Bowles & Gintis (2006) provided a comprehensive overview of what economists have termed "strong reciprocity", and how it can help understanding support for and opposition to welfare programs. The definition they give of strong reciprocity is:

> "a propensity to cooperate and share with others similarly disposed, even at personal cost, and a willingness to punish those who violate cooperative and other social norms, even when the punishing is personally costly and cannot be expected to entail net personal gains in the future." (Fong, Bowles & Gintis, 2006, p. 1441)

This definition corresponds with van Oorschot's (2000) survey and discussion of what he calls 'deservingness criteria', in which he concluded that reciprocity, colloquially put as "What have you done, or can you do, for us?", is one of the most important factors in determining if recipients of welfare fit the 'deservingness criteria' for approval. Alesina, Glaser & Sacerdote (2001) also discussed reciprocal altruism in their lengthy investigation into the differences in welfare support between the United States & Europe. They noted that anti-welfare forces emphasize welfare recipients such as the non-working poor as taking hard-earned money from tax-payers rather than working for a living to elicit resentment and hostility from those tax-payers through reciprocity norms (Alesina, Glaser & Sacerdote, 2001).

The existence of strong reciprocity norms as a primary motivator for support for welfare programs pokes several holes in traditional neoclassical explanations. Fong, Bowles & Gintis (2006) stated that:

> "Economists have misunderstood both the support for the welfare state and the revolt against welfare…, attributing the latter to selfishness by the electorate rather than the failure of many programs to tap powerful commitments to fairness and generosity and the fact that some programs appear to violate deeply held reciprocity norms." (p. 1460)

Another important collective factor is that of racial prejudice, which has also been cited by numerous researchers as an important contributary to opinion of welfare programs.

2.3.2 Racial Discrimination

Alesina, Glaser & Sacerdote (2001) noted that one of the primary strands of behavioural explanation for determining altruism, and through that channel support for redistribution, is racial prejudice. Pioneered by Becker's (1957) work on the topic, this literature notes that racial heterogeneity seems to be a significant factor in intra-community trust and social interaction, market outcomes, political processes, and interpersonal altruism (Alesina, Glaser & Sacerdote, 2001, p.227). They further detailed that historically racial animosity was used to defeat attempts at redistribution in the United States (Alesina, Glaser, & Sacerdote, 2001).

Meanwhile, Freeman (1986) asserted that migration by 'racially distinct' workers to Europe and the consequent increase in racial heterogeneity in the late 20th century resulted in similar factors destabilizing the status quo of strong welfare states; and came to the harsh conclusion that "national welfare states cannot coexist with the free movement of labor." While other researchers do not go this far, a breadth of empirical literature, such as Gilens (1995), confirms racial attitudes as a serious force in creating opposition to welfare.

Most relevantly to this thesis, racial priming is a key concept to explain how racial prejudice can be linked to and shape opinions of welfare programs. Mendelberg's (2001) framework contended that such priming must often be implicit and not explicit to function reliably, as otherwise it conflicts with equality norms. Given implicit priming, however, individuals' opinions and decisions

to support real policy can be significantly shifted (Mendelberg, 2001). In an empirical example of this, Harkness (2016) demonstrated discrimination in lending markets could occur based on the simple cue of manipulated avatar pictures in loan applications. Participants in Harkness' (2016) experiment showed significant deviations in lending decisions based on what she terms 'status expectations', though her results did also demonstrate an interplay between racial and gender prejudices in the execution of such decisions, with a surprising result: black female respondents were the second-most likely to have a loan approved.

The "race card" discussed by Mendelberg (2001) is another type of cue that can significantly alter opinions of policies (such as welfare programs), lending further evidence to the case shown in the motivated beliefs section that simple cues are capable of changing such opinions. That being said, that influence is often constrained and mediated by larger, omnipresent factors: institutions and culture.

2.3.3 Institutional & Cultural Factors

Institutional factors are widely considered to influence public support for government policy, particularly in literature discussing cross-national gaps in such support. The institutional comparative welfare-state literature historically relies heavily on Esping-Andersen's (1990) seminal welfare regime framework, which delineated three types of welfare regime – liberal, conservative & social democratic.

Subsequent literature with empirical backing by researchers like Papadakis & Bean (1993) asserting that "there is little support…that the popularity of welfare services is likely to vary with the institutional nature of welfare regime" (p. 246) generated further discussion and advances in this branch of the literature. Larsen (2008), while still referring back to Esping-Andersen's regime types, reframed how the institutional context of welfare regimes influences public support for welfare policy. He discussed three main characteristics as primary institutional influences on another factor: van Oorschot's deservingness criteria, which he identified as the channel through which institutional factors led to public opinion formation (Larsen, 2008). These three characteristics were: universalism in welfare policy itself, inequality in society, and the degree of job opportunities available to citizens (i.e. labour market structure) (Larsen, 2008).

According to Larsen, universalist welfare policy suppresses or modulates the other concerns discussed in this review (identity, reciprocity, need), as the policy comes to be perceived as a right rather than a privilege to be earned (2008). Reinforcing the neoclassical finding that rising inequality lowers support for redistribution (Moene & Wallerstein, 2003), Larsen (2008) noted that low inequality creates a 'negative' feedback loop that reinforces a status quo of support for welfare states—and conversely rising inequality can reduce that support as citizens begin to perceive consumers of welfare as different from themselves. Finally, Larsen (2008) argued that lower structural unemployment levels and the shift of wage-setting mechanisms to individual job-takers reinforces perceptions that the unemployed are responsible for their own unemployment, with predictably negative consequences for support for welfare states. Offe (1987) also referred to societal structure when considering the decline of the welfare state in modern western nations, arguing that, among other factors, the slow decomposition of self-conscious communities with a collective stake in welfare programs reflected a broader societal 'destructuration' that lead to a rapid loss of political support for the welfare state.

Jaeger (2009) introduced new empirical approaches to the welfare regime model, finding significant effects on public support for welfare programs by welfare regime type through the use of country-level regime indicators and Latent Profile Models. These were employed to control for the fact that the welfare regimes individuals experience may differ even within countries or cities (Jaeger, 2009). According to Jaeger, classifying entire countries as one regime type or another lacks the nuance necessary to detect such effects (2009). In a similar vein, Roosma, Gelissen & van Oorschot (2012) note that people's evaluation of welfare states is often multidimensional, and can vary: for instance responding positively to welfare policies' goals and range, but negatively when it comes to efficiency and effectiveness, particularly the underuse and abuse of benefits.

Contrasting Larsen's focus on labour markets and welfare regimes, Iida & Matsubayashi (2010) examined how constitutional designs might affect mass support for welfare policies. Conducting an empirical survey experiment, they found that individuals in countries with constitutions that clearly define welfare rights for citizens were more likely to have supportive attitudes towards welfare policies (Iida & Matsubayashi, 2010). The mechanism Iida & Matsubayashi (2010) suspected for this effect was that constitutional guarantees constrained elite discourse on the matter—elite discourse which would likely function as a cue to the rest of society on what 'correct' welfare opinions were, in a manner similar to Petersen et al.'s discussion of partisan cues.

The matter of religion's influence on other economic factors is almost always a thorny one, but it deserves a brief mention in this section. Stegmueller et al. (2012) advanced a persuasive argument that religious heterogeneity functions similarly to racial heterogeneity: the fact that religious individuals live in increasingly irreligious societies leads to an expectation that religious individuals will oppose income redistribution, something they found results supportive of for Catholic & Protestant populations in Europe after controlling for numerous other factors. They further noted that the important gap in support was not between different religious denominations, but between religious and irreligious respondents (Stegmueller et al., 2012). Alesina, Glaser & Sacerdote (2001) also provided evidence for this in their review, noting that while in the United States it comes behind race as a salient dividing line, in other parts of the world religious cleavages can be much more influential on the level of public support for welfare.

Analogous to an institutional factor, but often far more variable, is macroeconomic performance. Larsen's (2008) discussion of structural employment is also applicable in this arena, but more specific research has sought to investigate the direct link between macroeconomic crises and support for welfare spending.

2.3.4 Macroeconomic Factors

The consensus of the literature on how macroeconomic conditions influence public support for welfare and redistribution is decidedly mixed, and largely focuses on economic crises.

In a broad survey of how American opinion on welfare spending is affected by macroeconomic conditions, Kam & Nam (2007) tested survey data of individuals across almost three decades. They found that state-level inflation—but not state-level unemployment or productivity—predicted a shift in public support for specific means-tested welfare programs (those aimed at maximizing redistribution), though not support for the welfare state more generally.

Sihvo & Uusitalo (1995) found a different story in Finnish data in a similar timeframe, supporting the position that economic crises reduce public support for welfare spending, while it recovers in times of economic security.

Yet another interpretation comes from Soroka & Wlezien (2014), who draw on thirty years of British public opinion polling to emphasize a gradual, long-term shift against redistribution among the British public with no significant shift seen in response to the most salient economic crisis in recent times, the Great Recession of 2008.

Analogous to how Jaeger (2009) advances the discussion about the welfare state as an institution by introducing multidimensionality in measurement, Margalit (2013) uses a panel study across the years around the Great Recession in America to note that personal experience of hardship was a strong predictor of increased support for welfare spending. However, this effect was decidedly transient, and dissipated as individuals' employment situations improved.

The debate on how macroeconomic factors influence opinion on welfare spending is likely to continue, and its empirical focus will hopefully benefit from the renewed focus on multidimensionality seen in other strains of the literature, leading to more consistent results. Much more directly relevant to this thesis is the last literature topic discussed in this section, that of the deservingness heuristic:

2.3.5 The Deservingness Heuristic & Aarøe & Petersen (2014)

The deservingness heuristic is a critical component of Aarøe & Petersen's 2014 paper. Working from a basis in political science, they discussed several other explanations for differences in welfare support among different populations, such as the institutional path dependencies characterized by Esping-Andersen (1990), the racial heterogeneity argument discussed by Alesina, Glaser & Sacerdote (2001), and the religious and cultural values argument characterized by Stegmueller et al. (2012) (though they do not cite them directly). The thrust of Aarøe & Petersen's (2014) argument, however, focused on the micro-level influence of psychology on welfare support. Using the concept of a 'deservingness heuristic', they emphasized the possibility of a role for "differences in available information", i.e. media messaging and resulting opinion formation, in changing individuals' opinion on social welfare (Aarøe & Petersen, 2014, p.686).

This concept draws on much of the literature discussed in previous sections: the concept of strong reciprocity discussed by Fong, Bowles & Gintis (2006), the 'deservingness criteria' brought up by van Oorschot, the focus on dispositions in the case of complex altruism from Folbre & Goodin (2004), and the belief and opinion formation influences discussed by Benabou & Tirole (2016), Hawkins, Hoch & Meyers-Levy (2008) and Petersen et al. (2013). Aarøe & Petersen (2014) cited numerous further psychological studies that verify the operation of this heuristic as a regulator of altruism—precisely in the 'strong reciprocity' sense. It is from this concept and directly supportive literature in political sciences that Aarøe & Petersen (2014) drew their delineation of 'lazy' and 'unlucky' welfare recipients.

Aarøe & Petersen (2014) investigated this matter by conducting a cross-national experiment between Danish and American populations, arguing that despite the dissimilar stereotype predispositions of each nationality once individuals are exposed to 'deservingness cues', these differences melt away. To test this argument, they developed three measures to test for: a measure of default stereotypes about social welfare recipients per country, a measure of welfare

support under informational uncertainty about recipients, and a measure under informational certainty. The latter two correspond to the 'no cue' vs 'unlucky/lazy cue' conditions (Aarøe & Petersen, 2014).

Their social welfare stereotype measure was tested through a free-association exercise, while the informational certainty conditions were tested through a set of direct survey questions, whose exact wording is covered in the following section of this paper (Aarøe & Petersen, 2014). In summary, they created one control group of respondents who received no cues as to the deservingness of a hypothetical welfare recipient, and two treatment groups with opposing cues as to recipient deservingness (Aarøe & Petersen, 2014). Including demographic measures of gender, age & education, they had respondents rate to what extent welfare provision should be extended to their hypothetical recipient on a 1-to-7 Likert Scale ranging from 'Strongly Disagree' to 'Strongly Agree' (Aarøe & Petersen, 2014).

They took this experimental design to two approximately nationally representative 1000-person YouGov survey panels, one in the United States and the other in Denmark (Aarøe & Petersen, 2014).

The results for their first measure, the stereotype free-association, were reported as an independent samples t-test as well as a comparison of means (Aarøe & Petersen, 2014). They found that Americans had a significantly higher propensity to characterize social welfare recipients as lazy than Danes, and Danes a significantly higher propensity to characterize them as unlucky; in total, the American sample had a dominance of lazy stereotypes more than 4.6 times larger than that found in the Danish sample, matching prior literature (Aarøe & Petersen, 2014). With the difference in stereotypes established, Aarøe & Petersen (2014) moved on to a regression analysis of the responses to their survey, reporting significant differences between the American & Danish samples in the 'no cue' condition (where the different stereotypes presumably dominated) but no differences between the national samples in the cued conditions. They also interact the cue conditions with the measure of dominance of the lazy stereotype in each country to show that the cue conditions significantly crowd out (analogous to the concept discussed by Andreoni & Miller, 2002) the effects of national stereotype dispositions (Aarøe & Petersen, 2014). Their second regression analysis (Table 2 on p. 692) quite clearly shows a within-nation significant effect for each treatment group on opposition to social welfare depending on the cue conditions, but they do not discuss their results on within-nation terms, likely because the focus of the paper is on a cross-national comparison and the crowding out of national stereotypes (Aarøe & Petersen, 2014).

As this literature review has hopefully underlined, the determinants of public support for welfare programs are varied and multidimensional. Even a basic experimental contribution to this broad body of scholarly work can add to the validity of an explanation, and thus advance the state of knowledge in the field. With this overview complete, the next section proceeds to talk about the experimental design of this thesis.

3. Design

In order to test the replicability of Aarøe & Petersen's (2014) results on the influence of cues on welfare approval, I designed a straightforward experiment using a portion their basic experimental framework in a single population. As Aarøe & Petersen (2014) focus on a cross-cultural comparison, their experiment also includes a free-association portion on measuring default stereotypes that I do not. This precludes discussion of the validity of that portion of their paper. The portion I do investigate is that of the effect of deservingness cues, and whether these can be shown to have a replicable significant effect on the welfare opinions of individuals.

My design was conceived in advance of conducting the experiment, and both it and the hypotheses tested were pre-registered on the Open Science Framework (OSF) to preclude post-hoc experimental design. I conducted brief pilot tests of my survey with friends and family who were from a similar demographic as the subject pool (but not included in it) while developing the pre-analysis plan.

This section discusses my overall experimental design, from the research question investigated through the treatments, subject group, specific procedure and planned measured variables and statistical tests.

3.1 Research Question

In specific terms, the research question I seek to answer is: can the introduction of simple deservingness cues about a hypothetical welfare recipient spur differences in responses to a question about whether social welfare benefits should continue to go to recipients like the hypothetical person?

Concretely, this defines the focus of this thesis as replicating and testing a portion of Aarøe & Petersen's 2014 paper "Crowding Out Culture: Scandinavians and Americans Agree on Social Welfare in the Face of Deservingness Cues". While Aarøe & Petersen's paper focused on cross-cultural differences in stereotypes and the ability of cues to counteract the effects of these stereotypes; this paper differs in that it seeks to analyze the more fundamental question of whether such cues can significantly influence respondents' approval of social welfare benefits going to the person in the first place.

As the literature review suggests, there is a wealth of possible influences on individuals' approval of welfare programs. Confirming that deservingness cues can significantly shift that approval despite, or perhaps in concert with these other influences contributes to an understanding of the factors influencing welfare support and why it differs within and between populations.

3.2 Treatments

This study involved three treatment groups:

Treatment 0 – The Control or 'No Cue' Group

Treatment 1 – The Deserving or 'Unlucky Cue' Group

Treatment 2 – The Undeserving or 'Lazy Cue' Group

Treatment 0, the control group, was asked a question about approval of social welfare for an individual about whom they are given no deservingness cues, as follows:

> "Imagine a person who is on some form of social welfare. Should social welfare provision continue to benefit people like them?"

The response method available was a multiple-choice Likert scale, ranging from a value of 1 labelled as Strongly Disagree (No, It Should Not) to 7 for Strongly Agree (Yes, It Should).

The crucial factor for the control group was the lack of any cues about the person 'imagined' by the respondent. This would theoretically result in respondents revealing their aggregate opinion of welfare provision, combining the multifarious non-cue influences on this opinion.

Treatment 1, the 'unlucky cue' group, was given the following question:

> "Imagine a person who is on some form of social welfare. They have worked regularly in the past, but suffered a work-related injury. They have recovered and are very motivated to get back to work. Should social welfare provision continue to benefit people like them?"

The response method was the same as for the control group. As is hopefully apparent, this cue is meant to elicit an approving response based on the upholding of strong reciprocity norms as discussed by Fong, Bowles & Gintis (2006).

Treatment 2, the 'lazy cue' group, was given the following question:

> "Imagine a person who is on some form of social welfare. They are fit and healthy, but have not regularly held a job. They are not motivated to find work. Should social welfare provision continue to benefit people like them?"

The response method was once again the same. Similarly to the first treatment group, this cue was meant to elicit a response; however in this case it was a disapproving response based on the violation of those same reciprocity norms as discussed by Fong, Bowles & Gintis (2006).

A key point about these treatment groups, whose structure generally follows that of Aarøe & Petersen (2014), is that the wording of the cues was adapted somewhat from their original paper. Aarøe & Petersen's cues and questions read as follows (2014, p. 689):

> **Treatment 0**: "Imagine a man who is currently on social welfare.
>
> To what extent do you *dis*agree or agree that the eligibility requirements for social welfare should be tightened for persons like him?"
>
> **Treatment 1**: "Imagine a man who is currently on social welfare. He has always had a regular job, but has now been the victim of a work-related injury. He is very motivated to get back to work again.

To what extent do you *dis*agree or agree that the eligibility requirements for social welfare should be tightened for persons like him?"

Treatment 2: "Imagine a man who is currently on social welfare. He has never had a regular job, but he is fit and healthy. He is not motivated to get a job.

To what extent do you *dis*agree or agree that the eligibility requirements for social welfare should be tightened for persons like him?"

The changes made do not significantly alter the purpose of the cues used by Aarøe & Petersen and were made to try to address possible sources of error in the cue's effects to better determine whether such cues are significantly influential.

The first major change, from 'Imagine a man' to 'Imagine a person', was intended to strip out the possible influence of theoretical recipients' gender on respondents' answers. As Harkness (2016) shows, even small demographic cues can result in significant shifts in approval or disapproval for lending decisions; the same may be true for welfare benefit provision.

The second, moving from

"To what extent do you *dis*agree or agree that the eligibility requirements for social welfare should be tightened for persons like him?"

to

"Should social welfare provision continue to benefit people like them?"

sought to clarify the main question asked respondents, as Aarøe & Petersen's formulation seemed needlessly complicated and liable to confuse respondents. Saying 'should the eligibility requirements be tightened for persons like him' is euphemistic speech for 'should people like him continue to receive benefits'. I believe a more direct formulation would reduce possible bias or errors in respondents' answers.

The final major changes were in the treatment cues themselves.

In the first treatment, the fact that the motivated hypothetical recipient 'has recovered' was inserted to create more equivalence between the first and second treatments' hypothetical recipients. The implication of a work-related injury that was still affecting the first recipient, while the second was fit and healthy, added a possible health-care concern into an experimental question that primarily focused on an unemployment-based realm of welfare provision. This change was intended to have respondents in both treatment groups considering a largely identical social welfare situation, even if the precise social welfare type was left vague so as not to bring in possible bias by bringing specific real welfare programs to mind.

In the second treatment, the hypothetical recipient was depicted as not having regularly held a job, rather than the more extreme 'never had a regular job' formulation of Aarøe & Petersen. Aarøe & Petersen's formulation was presumably calculated to elicit a maximally negative response from strong reciprocity. However, it seemed possible it would bring mind a caricature rather than an actual hypothetical welfare recipient for respondents, thus reducing the consideration given to the question.

In the interest of having respondents think somewhat more about their response rather than simply selecting 'strongly disagree' and moving on I settled on the less strident formulation.

Of course, part of the adaptation of the treatment conditions was made with my subject population in mind, which is what I will proceed to detail next.

3.2 Subjects

Unlike Aarøe & Petersen (2014), who had access to two 1000-respondent panels of representative national surveys, my work uses the bread and butter of social science respondents: university students.

I investigated the possibility of surveying multiple university student populations, but due to time and resource constraints eventually settled on surveying specifically current students at the Stockholm School of Economics.

Conducting the experiment via an online Qualtrics survey, I sent a recruitment email out to a list of approximately 1700 current students across the university's student population of bachelors' and masters' students. This constitutes a significant portion of SSE's total enrollment, which is around 2000 students. Setting aside for now concerns regarding survey self-selection, this meant my responses would be at least generalizable to the SSE student population.

The survey was conducted in English on the assumption that students at SSE had a fluent command of English; SSE's master's and bachelor's programs are conducted entirely in English.

On the matter of exclusion criteria, I settled on a number of controls to try to ensure the subject population was not contaminated by multiple responses, unintended respondents, or unmotivated respondents:

1) **Non-completion:** An inherent exclusion was non-completion; as the experimental portion of the survey had only one question non-completion implied a lack of a response to record.
2) **Non-student participation:** To guarantee only students from my targeted subject population were responding, I required the submission of a valid SSE student email before the experimental portion. If something other than a valid SSE student email was submitted the response was discarded.
3) **Multiple participation:** While it is impossible to absolutely guarantee a lack of multiple participation as a result of students borrowing a friends' email, the above requirement for a valid SSE student email was also intended to limit participation. If the same SSE email appeared in the responses more than once, all of those responses would be discarded—a student 'gaming' the system to increase their chances of being chosen for the monetary reward would likely not be answering the questions seriously.

Privacy concerns prohibit the student emails as well as other personal details collected by Qualtrics (IP Addresses and locational data) from being publicly published in the data accompanying this thesis.

I did not implement a stopping rule for the collection of responses, as I had a clearly delineated survey period from the 24th of February to the 28th of March.

The target sample size for this survey was 300 responses; with the randomization option used this was expected to create control and treatment groups with approximately 100 respondents each. 100 respondents per group was the target to try to gather sufficient experimental power to support or reject the experimental hypothesis, as otherwise the results would be susceptible to an unacceptable likelihood of type M or S errors (Gelman & Carlin, 2014). With the survey sent out to around 1700 students, this assumed a response rate of approximately 17%, which based on prior experimental survey efforts at SSE was a reasonable measure. The introduction of a monetary reward component of 500 SEK to 6 randomly drawn respondents was intended to encourage participation, as this had an expected payout of approximately 10 SEK per respondent.

3.3 Experimental Procedure

To see the recruitment email, experimental survey and pre-registration plan in full, please refer to the Appendices (section 9), or consult the OSF listing for this experiment (linked in an earlier footnote).

On the 24th of February I sent out an email to a list of approximately 1700 students currently studying at the Stockholm School of Economics, containing a link to my experimental survey on the Qualtrics platform. Prior to encountering the experimental question, students were asked to sign a standard consent form, were informed of the possibility of winning a 500 SEK prize for participation, and were asked to report their SSE email in the survey itself for verification and prize-drawing. The e-mail to students informed them of the payout structure (a random draw of 6 email addresses, each receiving 500 SEK unconditionally), and it as well as the survey introduction attempted to dissuade discussion of the survey until it had run its course.

The initial pages of the survey read as follows:

[**Page 1**]

Welcome to the survey!

This is a short survey contributing to a Master's thesis at SSE. The expected time for completion will be about 2 minutes. Participation is anonymized - your responses will not be linked to your emails. The prize draw for **6 lots of 500 SEK** will be conducted from the email list by a third party. Please do not discuss the experiment with others after your participation until the response period has ended on the 28th of March, 2020. Your participation is very much appreciated!

[**Page 2**]

Online Survey Consent Form

Purpose of Research: To investigate economic decision-making

What you will do in this research: You'll participate in a decision-making study by answering some survey questions.

Time required: Participation will take approximately 2 minutes to fill out one (1) main survey question.

Risks: There are no known risks associated with this study. Your responses will be anonymized, as will your participation unless you are selected to win a prize.

Benefits: 6 participants will be selected randomly to receive 500 SEK if they complete the survey.

Anonymity: Your responses as well as your participation will be anonymous. The only exception is if you win the prize draw, in which case payout requires that your participation is known to the experimenter.

Participation and withdrawal: Participation is voluntary and you may quit the survey at any time.

To contact the researcher: This study is being conducted by Markus Jury from the Stockholm School of Economics. If you have questions about the research, please contact Markus at 41216@student.hhs.se

By selecting "I consent" below you indicate you have read and understood these conditions and agree to participate in the survey.

With Page 3 constituting of a prompt to enter their SSE Email.

Students were then randomly assigned to one of the three experimental groups through Qualtrics' question randomization option, which directed each respondent to a page containing one group's experimental question at random once they had confirmed their email submission.

The termination of data collection was planned to occur at the pre-specified ending date of the survey, regardless of how many responses have been collected. After the submission of my pre-analysis plan I considered breaching my delineated procedures to send out a reminder e-mail should the responses not reach my desired sample size, but thankfully this was not an issue.

3.4 Measured Variables

The primary outcome variable in this study was respondents' answers to the survey question, i.e. their degree of support for social welfare provision in the presence (or absence) of deservingness cues. Operationally, they chose one of seven options ranging from Strongly Disagree through to Strongly Agree on a Likert scale in the following format:

- ○ Strongly Disagree - No, It Should Not (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree - Yes, It Should (7)

One operation performed on this outcome variable is a simple summation of Likert scale values of a group's responses divided by the n of each group, to elicit the mean responses in each group for the purpose of descriptive statistics. In a similar manner, median as well as quartile responses are also collected, with the medians being particularly important for proper statistical testing as well.

Unlike Aarøe & Petersen (2014), this paper did not consider measuring numerous demographic variables. The reasons for this are: Firstly, their paper did not find any significant effects of these demographic controls on their results. Furthermore, the age and education variables were considered redundant, as the sample population consisted of a relatively small age range (likely with a small number of outliers) and were by definition all had a similar range of educational experience. Finally, including a lengthy extra page of demographic questions ran the risk of reducing the response rate, either due to student privacy concerns or simple boredom.

3.5 Statistical Testing

This paper further differs from Aarøe & Petersen in methodology: while the two researchers used OLS regressions across national groups to test the effects of their cues, this paper uses a Kruskal-Wallis H Test followed by Mann-Whitney U tests for pairwise comparisons. Without the motivation of various demographic controls to regress on, performing a direct statistical test was more appealing than running an OLS regression.

The Kruskal-Wallis H test is a non-parametric extension of the Mann-Whitney U test to more than two groups. Its null hypothesis is that sampled populations are identically distributed, i.e. that there are no statistically significant differences of an independent variable on a continuous or ordinal dependent variable between the groups. The alternative hypothesis being that at least one sample group is from a different distribution.

More specifically, a significant result in both the Kruskal-Wallis and Mann-Whitney tests indicates that (at least, in the case of Kruskal-Wallis) one sample stochastically dominates the other(s), meaning that there is a high probability that an observation from one group will be larger than an observation from another group.

In the case of this thesis, the Kruskal-Wallis H test tests whether approval of welfare benefit provision differs based on the deservingness cue shown to respondents. The null hypothesis is that there is no difference in approval for welfare benefit provision across all three groups, as the mean ranks of responses are not significantly different. The alternative is that there is a difference in approval, as the mean ranks are significantly different.

As it is a nonparametric test, it does not require an assumption of a normal distribution in responses from the population, which is something that can't be assumed with a behavioural response question. This is a key reason I chose the Kruskal-Wallis test, as I did not think my responses would be normally distributed.

In order to directly compare group medians it requires the various group distributions to have similar shapes and scales, but this is not necessary to draw conclusions from the test in general.

According to Siegel & Castellan (1988), there are three conditions to be able to use the Kruskal-Wallis H Test:

1. The dependent variable is continuous or ordinally dependent.
2. There are at least two independent groups in the independent variable.
3. Observational independence: different observations within groups are unrelated.

The first two conditions are definitely fulfilled by this experiment, and there are no reasons to expect an unusual lack of independence between observations.

If the Kruskal-Wallis H Test returns a significant ($p<0.005$) result in rejecting the null hypothesis, each groups' distribution will be compared with each of the other groups using pairwise Mann-Whitney U tests. Should the test return an indicative but not a significant result ($p<0.05$), the Mann-Whitney U tests will proceed with their results considered as suggestive evidence.

As discussed above, the results of tests will be treated as significant if they are under the p-value threshold of 0.005. If the results show this, it means that the distribution of responses to questions differed between groups. If the groups responses' also have similarly shaped distributions, it would also show that the group medians are significantly different. They will be treated as suggestive evidence if they are under the more traditional p-value threshold of 0.05. Should the results not reach either value, it means that there was no significant difference in the distribution of the groups' responses.

The Kruskal-Wallis H Test & Mann-Whitney U pairwise tests to be conducted are two-tailed.

4. Expected Results and Limitations

This section deals with the hypotheses forwarded by this thesis, what I expected the results of those hypotheses and the experiment at large to be, and a discussion of experimental limitations faced by the experimental design.

4.1 Hypotheses

As discussed briefly in the Design section, I perform a Kruskal-Wallis H Test to test my main research question of whether there are significant differences in welfare provision approval based on deservingness cues. This generates the following hypotheses:

H1: *Null:* There is no systematic difference in approval for welfare provision depending on deservingness cues presented to respondents.
Alternative: There is a systematic difference in approval for welfare provision depending on deservingness cues presented to respondents.

Assuming the alternative hypothesis of this test is accepted at either a significant or suggestive level, the following Mann-Whitney U hypotheses would be tested:

H2: *Null:* There is no systematic difference in approval for welfare provision between the 'no cue' and the 'unlucky cue' group.
Alternative: There is a systematic difference in approval for welfare provision between the 'no cue' and the 'unlucky cue' group.

H3: *Null:* There is no systematic difference in approval for welfare provision between the 'no cue' and the 'lazy cue' group.
Alternative: There is a systematic difference in approval for welfare provision between the 'no cue' and the 'lazy cue' group.

H4: *Null:* There is no systematic difference in approval for welfare provision between the 'unlucky cue' and the 'lazy cue' group.
Alternative: There is a systematic difference in approval for welfare provision between the 'unlucky cue' and the 'lazy cue' group.

Based on the results found in Aarøe & Petersen (2014) as well as the strong reciprocity literature detailed by among others Fong, Bowles & Gintis (2006), I had certain expectations about each of these hypotheses:

For **H1**, I expected that there would be a significant result and thus a systematic difference in registered approval between the three groups. More specifically, I expected at least some group to be stochastically dominant over at least some of the others.

For **H2, H3 & H4**, I expected each to return a significant result and thus a systematic difference between the two groups tested in each hypothesis. When it came to the direction of these results, I expected the first treatment group ('unlucky cue') to be stochastically dominant over each of the other two groups, and the control group ('no cue') to be stochastically dominant over the second treatment group ('lazy cue').

I also expected to be able to speak even more specifically about a difference in group medians (following a similar pattern to the above discussion of stochastic dominance between groups), but as I will discuss in the analysis section limitations of my experimental design precluded drawing such conclusions.

I expected a sample size between 350 and 500, as survey efforts using the same procedure at the Stockholm School of Economics within the past year had achieved sample sizes around those numbers.

4.2 Experimental Limitations

As with any experimental thesis, I faced limitations that circumscribed the applicability and reliability of my results. The implications of these are discussed further under the internal and external validity sections of the discussion; this section is here to list and briefly expand on those limitations. Some of these were inherent to the experimental design, while others were due to unexpected results or mistakes in following said design.

As an online survey experiment, the usual caveats regarding such experiments apply. I could not control in what conditions—physical or mental—respondents took the survey and to what extent these influenced responses. I was, however, able to see the amount of time respondents took to answer the survey due to Qualtrics' data collection. I cannot be certain that students did not use other students' emails to enable multiple participation, nor can I be certain that they did not have a non-student complete the survey for them.

I cannot rule out that the mistake I made regarding the accessibility of the survey at the very beginning of the survey period created a bias towards less eager student respondents by causing the attrition of at least some of the students who opened the link in that time.

The expected effect size of the deservingness cues was difficult to quantify, as Aarøe & Petersen (2014) perform their analysis with a regression on the respondents' country of origin and do not report a group or pairwise analysis of the basic experimental conditions' (deservingness cues) effects. My target survey size was chosen based on a balance of what could generate sufficient experimental power while remaining realistic for the survey method employed; but it could still be too low to avoid type M or type S errors as discussed by Gelman & Carlin (2014), or even a simple Type II error. This is a particular risk as the paper I am working from had sample sizes over three times as large as mine.

5. Results and Analysis

This section details the experimental results from gathered responses and statistical testing, as laid out in the pre-analysis plan. It then proceeds to analyze these results in the context of the design of the experiment. Some minor corrections were made from the pre-analysis plan to the interpretation of the hypothesis testing results, especially in light of distributional concerns raised after doing descriptive statistics, but the analysis plan itself is still adhered to.

As a quick summary, the results show that there is a significant difference between the 'lazy cue' group and the other two, as evidenced by a significant Kruskal-Wallis H Test result for **H1**, followed by significant values in the pair-wise comparisons for **H3** and **H4**. I am able to reject the null hypothesis for each of these tests. The results for **H2** found no significant difference in approval for welfare provision between the 'unlucky cue' and 'no cue' conditions, meaning I am unable to reject the null in that case.

5.1 Statistical Reporting

The survey ran from the 24th of February to the 28th of March, a period that totaled 4 weeks and 5 days. During that period, 348 students accessed the Qualtrics survey. Of those students, 324 completed the survey; 24 responses (6.9%) were left incomplete. Of these incomplete responses, 13 stopped after agreeing to the consent form but before filling in an email, and 11 stopped after filling in an email but before being shown the experimental question. As none of these students were assigned to an experimental group, there is no need to worry about attritional bias.

Of the valid responses to the survey, 276, or 85%, were submitted within a day after the email was sent on the 24th of February. 309, or 95%, came within the first 4 days. 231, or 71% came within *5 hours* of the email being sent out. The duration of time spent on the survey ranged from a minimum of 22 seconds to a maximum of 4026 seconds, with a median of 66 seconds and only 13 responses taking longer than 4 minutes.

I downloaded the data on these responses from Qualtrics in the form of .csv file, then saved in .xls form for importation into R.

Before importing the data into R, the 324 responses were further trimmed to 320 due to 3 responses lacking a valid email submission, and one turning out to be an alumnus of SSE and thus outside of the strict subject population. Five responses had invalid emails only due to mild typographical errors (e.g. 00000@studnet.hhs.se instead of 00000@student.hhs.se). These were corrected and included in the response pool. I then removed superfluous information provided by Qualtrics: respondent IP addresses, geolocational data, response and completion status/progress, distribution channel type, user language and response form (online or not).

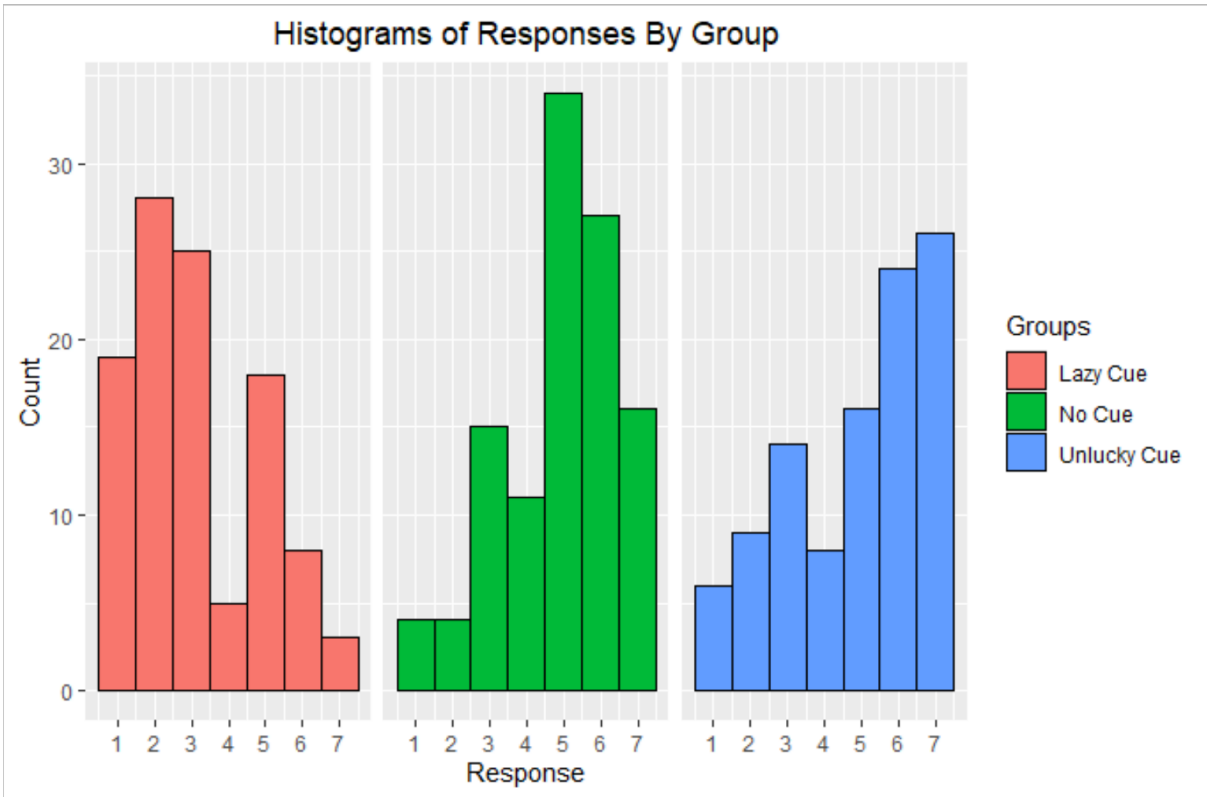
Proceeding to R, I labelled the data and after removing data irrelevant to statistical testing (date and time of responses as well as their unique Qualtrics identifier) proceeded with summary statistics, listing the group means, standard deviations, minimums and maximums, medians, and interquartile ranges:

Table 1: Response Summary Stats by Treatment

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
No Cue	4.910	1.541	1.000	4.000	5.000	6.000	7.000
Unlucky Cue	4.893	1.899	1.000	3.000	5.000	6.500	7.000
Lazy Cue	3.104	1.690	1.000	2.000	3.000	5.000	7.000

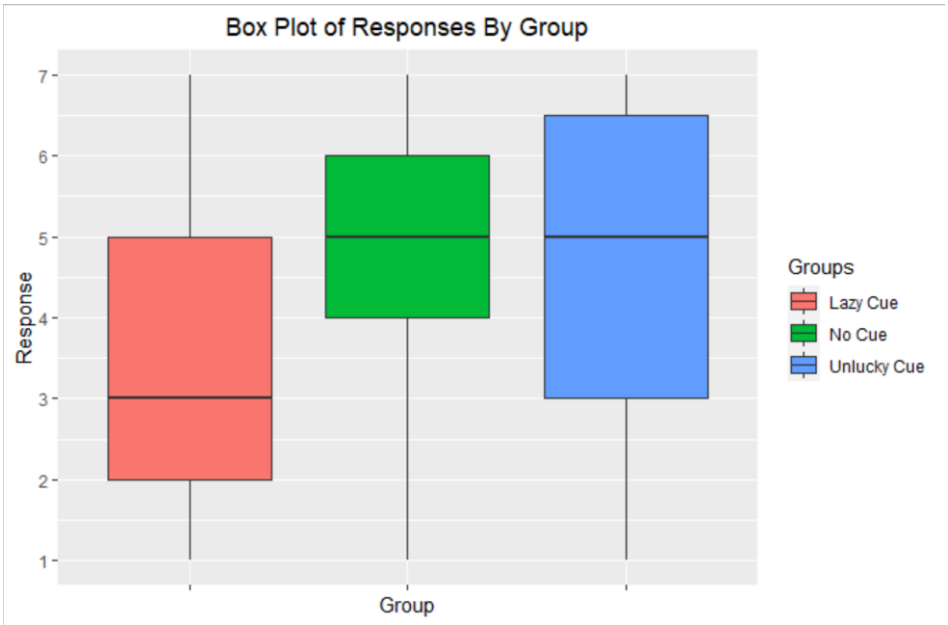
I also created basic histograms of the distribution of responses in each group for comparative purposes and to check assumptions related to the Kruskal-Wallis and Mann-Whitney U tests I would be performing:

Figure 1: Histograms of Responses by Group



Finally for the descriptive statistical portion, I created a box plot to further investigate the distributional properties of each experimental group:

Figure 2: Box Plot of Responses by Group



After this, it was time to perform the planned statistical testing. The Kruskal-Wallis H test results were as follows:

Table 2: Kruskal-Wallis H Test Results

	Results
Kruskal-Wallis chi-squared	62.234
df	2
p-value	3.06e-14

The results rejected with statistical significance the null hypothesis **H1**: there is a systematic difference in approval for welfare provision depending on deservingness cues presented to respondents.

This statistically significant result ($p<0.005$) meant I could proceed with the second portion of my statistical testing analysis plan, the pairwise Mann-Whitney U tests. The results for those were as follows:

Table 3: Mann-Whitney U Test Results

	Unlucky Cue	Lazy Cue
Lazy Cue	$p<1.485e-10$	
No Cue	$p<0.583$	$p<2.17e-12$

Here the results were more mixed. The results for **H3** and **H4** allow for rejection of their respective null hypotheses: there is a systematic difference in approval for welfare provision between the 'no cue' and the 'lazy cue' groups ($p<0.005$), and a systematic difference in approval for welfare provision between the 'unlucky cue' and 'lazy' cue group ($p<0.005$). The results for **H2**, however,

were not significant at any level ($p>0.05$), meaning I could not reject the null hypothesis that there was no systematic difference in approval for welfare provision between the 'no cue' and 'unlucky cue' groups.

5.2 Analysis

The response rate for the survey in general was somewhat below expectations, as prior surveys of a similar length at SSE targeting a similar number of students had achieved sample sizes above 400. That being said, a number of factors may have lowered the response rate of students. The first is that only one e-mail was ever sent out to students about the experiment. The extreme tendency for respondents to respond soon after receiving the email suggests a 'reminder' email could have significantly boosted the sample size. As it was not planned for in the pre-analysis plan, however, I chose not to pursue this.

Another factor that may have limited response rate, especially considering the extreme tendency of responses to come soon after the email had been sent, is the fact that for approximately 5 minutes after the email was sent out the Qualtrics survey was not accessible. Due to a mishap involving the time zone Qualtrics was operating out of, the survey was still 'locked' when the email went out, and this was only corrected upon receiving two emails from confused students. While the response rate was still high enough to reach the target sample size, it is likely this caused at least some attrition as the most responsive students opened the email link to find a locked survey and thereafter discarded the email without checking again.

The response times registered were about in line with the expected response times from pilot survey testing (generally 40-120 seconds), barring a few extremely lengthy outliers that were most likely due to respondents being distracted mid-survey and only coming back to it a while later. While the very fastest responses were generally around 30 seconds, these were not so fast that I worry they constituted respondents clicking through the survey mindlessly.

Proceeding to an analysis of the summary statistics, every cue's responses spanning the full range of the provided Likert scale is essentially expected; on this front the histograms and box plots provide a more meaningful descriptive result. We can however already see the large discrepancy between the means and medians of the first two groups ('no cue' and 'unlucky cue') and the last treatment groups ('lazy cue'): 'no cue' and 'unlucky cue' are extremely close in both these measures, while the 'lazy cue' results lag behind noticeably. This is not matched by their standard deviations, which are reasonably close to one another, suggesting that heteroscedasticity does not endanger my results. The one measure in which the groups measured by **H2** noticeably diverge is in their 25^{th} percentile results, in which the 'unlucky cue' group is as far from the 'no cue' group as it is the 'lazy cue' group. Given the later statistical testing results, however, this cannot be said to amount to a significant difference.

As we can see from the histograms, the existence of low-approval responses for the 'no cue' and 'unlucky cue' groups was limited, as was the existence of high-approval response for the 'lazy cue' group. The most extremely contrarian responses (1s for the first two groups, 7s for the 'lazy cue' group) made up less than 6% of total responses for the first two groups, and less than 3% of total responses for the 'lazy cue' group. This suggests the minimum and maximum measures are of limited use in understanding the data. The histograms also show an interesting pattern, in which respondents avoided settling for the neutral 'neither agree nor disagree' option, even in the condition with no cue given. A distinct negative skew applies to the 'no cue' and 'unlucky cue'

groups, whereas the ‘lazy cue’ groups’ responses display a distinct positive skew that approaches the inverse of the other two. This difference in distributional shape is very important when it comes to interpreting the results of the statistical testing performed: as I cannot assume each of the groups’ distributions have the same shape and scale, I cannot discuss the results in terms of group medians, due to that being the necessary assumption to do so with the Kruskal-Wallis H & Mann-Whitney U tests.

This difference in distributional shape is also apparent in the box plot graph, where the ‘lazy cue’ group’s box is shifted strongly downward from the other two groups’ boxes. The medians of the ‘no cue’ and ‘unlucky cue’ groups being equal is more visible than in the histograms.

Moving on to the results of the Kruskal-Wallis H test, the chi-squared H value of 62.234 with 2 degrees of freedom is strongly statistically significant, resulting in a very low p-value. While this emphatic result does give me confidence that a cue effect exists, it unfortunately says nothing about the possible effect size or reliability of the result. To do so quantitatively I would have to perform further effect size tests, which I did not plan for in my pre-analysis plan and thus leave for further exploratory study outside of this thesis. That being said, the markedly obvious difference in distributional shapes shown in the box plots and histograms suggests a relatively strong effect size.

The Mann-Whitney U pairwise results provide a similar statistically significant result in 2 out of 3 pairwise tests conducted. I can conclude that the introduction of a negative ‘lazy cue’ has a significant effect on approval for welfare provision but cannot reject the null that a positive ‘unlucky cue’ has no effect on approval for welfare provision. As I will discuss further in the next section, this discrepancy could be for a number of reasons, and it may be that either my subject pool or experimental design functioned to conceal the effect of the ‘unlucky’ cue.

6. Discussion

This experimental study studied whether simple deservingness cues can create differences in responses to a question about whether social welfare benefits should continue to go to a hypothetical recipient. Modelling the questions and procedure on that of Aarøe & Petersen (2014), I show that such deservingness cues can create significant differences in responses even among a single homogenous population such as a body of students from the same university. Specifically, I reject the null hypotheses that the 'lazy cue' group responses are not systematically different from the responses of the 'no cue' group and the responses of the 'unlucky cue' group. However, I fail to replicate the effect of a 'deserving cue' on responses shown by Aarøe & Petersen (2014) and will discuss in the interpretation section why this may have occurred.

This constitutes a contribution to the body of work on the topic of deservingness cues, and on the usefulness of the deservingness heuristic for explaining welfare state approval more generally.

In light of these results, I proceed to discuss the implications of my results in terms of the theoretical literature, with consideration toward the unique characteristics of my subject population. I will then conclude by discussing the experimental limitations I faced, and their impact on the experiment's internal and external validity.

6.1 Interpretation of Results

The rejection of the null hypotheses of **H1, H3** and **H4** replicate the results of Aarøe & Petersen (2014) and could carry a similar interpretation. That interpretation being that the introduction of the 'lazy cue' disrupts preexisting dispositions and beliefs by engaging the concern for deservingness noted by van Oorschot (2000). The strong reciprocity hypotheses discussed by Fong, Bowles & Gintis (2006) as well as Alesina, Glaser & Sacerdote (2001) also lend credence to this explanation: faced with considering a hypothetical recipient who violated expected reciprocity norms, respondents shifted their positions on the matter of social welfare for that recipient. The conclusion that simple cue sentence added on to the introduction of a hypothetical recipient is able to create such a shift is bolstered by the findings of Hawkins, Hoch & Meyers-Levy (2001) as well as Petersen et al. (2013) and Krueger & Rothbart (1988).

What exactly contributed to the evolution of the norms that created such a reaction—in the sense of norm evolution discussed by Ostrom (2001) and Young (2015)—is beyond the scope of this paper to test experimentally, but I will discuss a few possible explanations with reference to the literature and my subject pool. It is likely the norms generating such a reaction reflected upper-middle class Swedish (or at least European WEIRD) society more generally. While there is some demographic heterogeneity within the student body at the Stockholm School of Economics, it is still overwhelmingly WEIRD, and largely hails from a middle-to-upper-class income bracket. At first glance, this would suggest it is unlikely that at least some of the channels for the transmission of pro-universal-welfare norms identified by Barón, Cobb-Clark & Erkal (2015) would be missing in my subject population, particularly a lack of experience of receiving welfare.

However, this could well be counteracted by the existence in Sweden of an easily-accessible monthly study allowance of 1,250 SEK for students up to the age of 20, which is mostly universal—eligible are Swedish citizens as well as foreigners with permanent residency—and widely utilized. The existence of this stipend means it is likely a significant portion of my subject population has

direct experience in benefiting from a welfare program. As this benefit has existed since the introduction of the bills establishing it in the years 1999 and 2000, it is safe to assume that all of the Swedish citizens in my sample had the ability to take advantage of the benefit. As for what proportion of my sample that might be, a recent report by the Stockholm Academic Forum notes that SSE has the highest proportion of international students in Stockholm, with 30% of incoming students consisting of international arrivals (Lundström, 2020). Regardless, this still means that at least 70% of my sample either has or has had direct access to or experience with a generous universal welfare program—and the rest likely have a friend or two who do.

The existence of this stipend may help explain why I could not reject the null hypothesis of no systematic difference between the 'no cue' and 'unlucky cue' conditions. The universality and direct experience of at least the existence of the stipend on the part of my subject population follows one of Larsen's (2008) examples of the institutional factors that generate support for welfare programs. This may have shifted the respondents' default view of welfare recipients to consisting of people similar to themselves, who would—in line with Benabou & Tirole's (2016) motivated belief framework's concept of identity-enhancing behaviours—proceed to approve of others receiving similar benefits. Thus a conflation of the 'no cue' and 'unlucky cue' response groups: in the absence of informational cues, students with a high likelihood of positive experiences with welfare programs may assume a hypothetical recipient is, like themselves or someone they know personally, a 'deserving' recipient. Bringing in Iida & Matsubayashi's (2010) constitutional discussion, Chapter 1, Article 2 of the Swedish Constitution asserts:

> "The personal, economic and cultural welfare of the individual shall be fundamental aims of public activity. In particular, the public institutions shall secure the right to employment, housing and education, and shall promote social care and social security, as well as favourable conditions for good health." (Sveriges Riksdag, 2016, p. 65)

As this is a strong definition of welfare rights, if Iida & Matsubayashi's (2010) explanation holds merit it is likely it constrained the exposure of a majority of my subject population to an elite discourse disparaging the welfare state. Combined with Petersen et al.'s (2013) discussion of how political cues may shift public opinion formation, a lack of negative cues in politics—which is generally dominated by elite discourses—could well have influenced even students who are only somewhat attentive to politics.

The obvious reaction to a default position of seeing welfare recipients as similar to oneself and therefore 'deserving' comes when this similarity is challenged by the apparent violation of a deeply-held social norm such as strong reciprocity. At this point motivated beliefs likely work in the opposite direction and strengthen the rejection and exclusion of the offending party. This seems likely to be part of the explanation as to why I saw such a stark cleavage in response distribution shape between the 'lazy cue' group's responses and those of the other two groups. Krueger & Rothbart's (1988) work on the influence of trait terms overriding 'category' judgements also supports this interpretation.

Moving on to less important explanatory strands for the topic of this thesis, neoclassical economic explanations for support for redistribution do not offer answers to most of my interpretation. However, students will generally speaking be below the mean income, and going by models in the vein of Meltzer & Richard (1981) this would imply their default stance would be approving of increasing redistribution generally. I do not have enough information about my student population to conclusively speak to the explanatory power of the racial heterogeneity argument

on their pre-existing dispositions toward welfare programs. I cannot rule out that the 'lazy cue' functioned for at least some respondents as an implicit 'race card' (per Mendelberg, 2001) and that this influenced their evaluation of the hypothetical recipient, but considering the simplicity and academic context of the cue I do find it unlikely. While my subject population likely all had first-hand experience of an economic crisis through the 2008 Great Recession, the mixed results of such crises on welfare support in the literature and especially Margalit's (2013) affirmation that any such opinion shifts are transient makes this an unlikely to have influenced responses in my samples.

In conclusion, I was able to confirm through my experimental results that at least some simple deservingness cues can create a systematic difference in responses about—and thus presumably opinions of—the provision of welfare to hypothetical recipients. This partially confirms Aarøe & Petersen's (2014) broader results on the ability of such cues to 'crowd out' inherent stereotypes, and sample conditions as well as experimental limitations seem likely to have influenced the failure to replicate a systematic difference between the 'no-cue' and 'unlucky cue' conditions. With this in mind, I now proceed to a discussion of my experimental limitations.

6.2 Discussion of Limitations

With this in mind, I now proceed to a discussion of my experimental limitations.

The first limitations that come to mind regarding this experiment are the usual cohort of limitations that accompanies an online survey method: a lack of control over experimental conditions, difficulty being absolutely certain participation is limited to the subject group, and subjects having problems with accessing or using the survey tool at a distance. I consider these and other concerns related to online survey completion real but minor issues, as a large body of scholarly work such as that by Amir, Rand & Kobi Gal (2012) has shown online survey methods to be comparable to in-person experimental methods in reliability.

On the topic of the survey being locked for the first five minutes following the recruitment email: despite the exceedingly short amount of time this was an issue, it could still have biased my sample away from students who were most eager to take part in such an experiment. However, I feel those students would also be those most likely to check the email again at a later point to see if it had been fixed; this supposition is supported by the fact that I received three separate emails from such students asking why the survey was still locked. While I certainly regret this error, as such I do not think it undermines the validity of my results.

When it comes to a lack of expected effect sizes, the supporting literature ameliorates this concern somewhat. There is an extensive literature supporting deservingness heuristics playing a large role in determining public opinion of welfare (Fong, Bowles & Gintis, 2006), so I expected the true effect size would still be large enough to provide useful results with my significantly smaller sample size. This certainly proved to be the case in the 'lazy cue' condition. Though it may have contributed to a lack of systematic differences between the 'no cue' and 'unlucky cue' conditions and a larger sample size of SSE students could have uncovered some level of difference between respondent distributions, I believe a good portion of the lack of difference could be accounted for by the countervailing largely belief-based factors discussed in the previous section. Nevertheless, the effectiveness of a "deserving recipient" cue similar to my 'unlucky cue' would

be a fruitful area for further research, perhaps with a student population in a country less inclined towards the 'unlucky' position by default.

With the aforementioned limitations covered, I now move on to a discussion of the broader validity of my findings.

6.3 Discussion of Validity

A major concern for any scientific work is that of internal and external validity.

6.3.1 Internal Validity

There are a number of factors that strengthen the internal validity of this experiment. The Qualtrics survey tool truly randomized—as much as any computer can—the assignment of respondents to experimental groups, ensuring proper randomization. A pre-registration clearly laid out my process of analysis and was followed throughout the experiment, limiting the scope for researcher degrees of freedom on my part. While I did suffer an apparent attrition rate of approximately 7%, with the survey having only one experimental question this form of attrition is unlikely to have lead to any form of survivorship bias. It is hard to imagine any serious confounding variables that could have influenced the results of my experiment, as it was very streamlined to begin with. The partial confirmation of Aarøe & Petersen's (2014) results by mine that are mirrored in the wider deservingness heuristic and cue literature such as Krueger & Rothbart (1988) also suggests the fundamental composition of my experiment was sound.

However, there are certainly some concerns for internal validity I must address.

As discussed in the experimental limitations section, the online survey nature of this experiment as opposed to it being an in-person laboratory experiment comes with a variety of internal validity concerns that, while perhaps minor, could still add up to undermine the validity of my results. While the literature suggests this is unlikely to be the case, conducting a similar experiment in a laboratory format would be a good test to see if these concerns cumulatively constitute a real issue.

As with any survey experiment that actively recruits from a larger population, self-selection bias is a concern for the validity of the experiment. In the case of this experiment the concern is not about demographic self-selection. Aarøe & Petersen's (2014) work indicates that self-reported gender has little influence on the results, and the population surveyed was largely homogenous on the variables of education and age. The concern is particularly belief-based self-selection. It seems quite possible that the students most interested in participating in an experimental survey share a tendency toward generosity, altruism & strong reciprocity ('I take this experimental survey to help your thesis, and in the future someone will help me out'). If this is the case, it may have fundamentally shifted my respondent population away from the general student population at SSE and toward those students most likely to cause, for instance, a shift of the 'no cue' group toward the same responses as the 'unlucky cue' group. In my case I had few other options when it came to recruitment, but future research could employ a different recruiting method to see the extent of bias that self-selection causes in this case.

Another challenge to the internal validity of this experiment is a possible flaw in the design of the control group. The 'no cue' condition is supposed to elicit an individual's aggregate opinion of

welfare provision without cues, but this is not something I could clearly test for while retaining a short survey that would be most likely to reach my desired sample size. An experiment with more compelling reasons to prevent attrition due to fatigue (such as a compensated laboratory experiment format) could incorporate questions about subjects' general welfare opinions to attempt to compensate for this; though that also runs into issues regarding the timing of those questions and their effects on the experiment proper.

Much future research will surely be written about the effects of the Coronavirus crisis on almost everything in human society, including approval for welfare programs. Thankfully for the internal validity of my results, my survey period occurred before the effects of the epidemic reached Europe, so this extraordinary event will have had little bearing on my subjects' responses.

A final concern for internal validity that I consider separately from the more usual concerns with online surveys is the possibility of subjects in my experiment talking about it amongst themselves. As I drew from a rather limited subject pool, many of whom know each other, I cannot rule out that my results included some level of subjects talking about it amongst themselves or responding in each other's company. This is another validity concern that was difficult to avoid due to the subject pool constraints I faced, and could be ameliorated by conducting the experiment in a different manner such as an in-person laboratory format.

6.3.2 External Validity

The external validity of this experimental study is limited, of course, by any of the factors mentioned in the internal validity section that might undermine its general results. Beyond that, however, there are several further challenges to its external validity.

The first and most obvious is that my subject pool was in no way generalizable to the general public, even within Sweden. The Stockholm School of Economics is a prestigious research university that focuses on business and economics majors, further restricting the generalizability of my results to a subset of western European university students. As with so many social science experiments conducted at universities, my sample was decidedly WEIRD, with all the limitations to external validity that entails (see Henrich, Heine & Norenzayan, 2010). While the diversity of the Stockholm student population & by extension SSE continues to grow, the five largest countries of origin for international students are Italy, Spain, Greece, the USA, and the UK—so still weird (Lundström, 2020).

Another limit to the generalizability of my results to the broader sweep of the 'public support for welfare' debate is the issue of survey responses to my explicitly academic survey not functioning as a proper representation of real-world opinion formation. Even if within my experiment the cues function to change subjects' responses to a hypothetical scenario, it does not naturally follow that this behaviour is mirrored in non-academic real-life situations. The lengthy literature on the influence of cues in the political (Petersen et al., 2013) and economic (Hawkins, Hoch, & Meyers-Levy, 2001) spheres where the academic nature of the experiment was not immediately apparent does somewhat ameliorate this concern, but it remains one to be mindful of. A possible further step to address these concerns would be to devise a field experiment to test the influence of cues outside the lab; however it is likely such an experiment would have difficulty with experimental ethics concerns.

Another limitation in this vein concerns the concept of multidimensionality introduced by Roosma, Gelissen & van Oorschot (2013): it is possible that my questions are eliciting opinion about a specific part of welfare provision that is not generalizable to how SSE students' opinions of various other parts of it would respond to deservingness cues. Similarly to the earlier discussion on the internal validity of the 'no cue' condition, a more expansive survey that manages to capture subjects' opinions on welfare topics might shed light on this topic.

Despite the limitations and challenges to validity faced by my experiment, I am still confident this study manages to contribute to the broader literature on deservingness cues

7. Conclusions

The welfare state is a major component of modern society: public social spending across the OECD averages a massive 26% of GDP (Barr, 2020). As noted in much of the literature surveyed, the extent of such programs and more particularly for this paper the support for their implementation and continuation varies significantly between and within populations across the globe. The importance of such support for the continued existence of such programs cannot be understated, as a wide body of scholarly work in the past several decades attests (Larsen, 2008; Offe, 1987).

In this experimental paper I designed an online survey experiment to examine a specific sub-topic of the public support for welfare literature, that of perceived recipient deservingness as a motivator for or against such support. Basing my design on an experiment conducted by Aarøe & Petersen (2014), I adapted their design to suit my more limited experimental means and found results that replicated their result that deservingness cues are able to create a significant difference in responses to questions about the provision of welfare to hypothetical recipients. Across three treatments, my pairwise results showed these differences were present when comparing a question with no cue to one with a cue signalling undeservingness, as well as when comparing a question with a cue signalling deservingness with one signalling undeservingness. I could not reject the null hypothesis of no difference in responses between a question with no cue and one with a cue signalling deservingness, however. My findings suggest that SSE students are by default inclined to view welfare recipients as deserving, but that they exhibit a marked level of strong reciprocity, which is also termed the 'deservingness heuristic' in the literature.

Previous literature is expansive on suggesting possible motivations behind public opinion on welfare spending, redistribution, & the welfare state. Such opinions are inevitably multidimensional (Roosma, Gelissen & van Oorschot, 2013), making measurement and the generalizability of any one theory difficult. Nevertheless, many different academic disciplines have delved into this topic, ranging from neoclassical economics, through behaviour economics and psychology, and proving most prevalent in the fields of political science, sociology and political economy. Each of these fields has made important contributions to the topic, without which the state of knowledge on the topic would be poorer.

My specific findings sought to replicate a subset of Aarøe & Petersen's (2014) results. While not extremely ambitious, I believe my findings contribute to the state of knowledge in the field. The lack of rejection of the null in the case of the 'no cue'-'unlucky cue' comparison may indicate that SSE students as a cohort hold a decidedly positive default opinion of welfare recipients, which is only soured by apparent violations of strong reciprocity.

My research provides several possible directions for further research: in-person or field experiments on the same topic could yield important information on whether online textual cues differ from verbal or non-electronic cues. A survey also eliciting information about subjects' dispositions towards various welfare provision topics might be able to isolate which specific predispositions deservingness cues manipulate or 'crowd out', similar to Aarøe & Petersen's (2014) work with one such aspect in stereotypes. A randomized recruiting method that did not rely so much on self-selection could deal with concerns that results on deservingness cues are skewed by the generous predispositions of self-selected subjects.

Finally, I believe that on a broader level my thesis, and the topic of deservingness cues more generally, has important implications for how societies approach messaging around welfare topics. With research indicating ambient societal cues can influence opinion formation in the realm of politics and economic consumption (Hawkins, Hoch & Meyers-Levy, 2001; Petersen et al., 2013) it is likely that public discourse in politics and the media influencing opinion formation about welfare states. Economic policymaking is often constrained by the realm of public opinion. Further knowledge of when and how such cues are transmitted (and such opinion formation occurs) may be crucial to preventing the hijacking of public opinion by special interest groups like those discussed by Tullock (1970) to turn it against otherwise sound redistributive policy.

8. References

Aarøe, L., & Petersen, M.B. (2014). Crowding out culture: Scandinavians and Americans agree on social welfare in the face of deservingness cues. *The Journal of Politics, 76*(3), 684-697.

Amir, O., Rand, D.G., & Kobi Gal, Y. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE, 7*(2), e31461.

Andreoni, J. (1989). Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy, 97*(6), 1447-1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal, 100*(401), 464-477.

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*(2), 737-753.

Akerlof, G.A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488-500.

Alesina, A., Glaeser, E., & Sacerdote, B. (2001). Why doesn't the United States have a European-style welfare state? *Brookings Papers on Economic Activity, Vol. 2001*(2), 187-254.

Arrow, K.J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review, 53*(5), 941-973.

Atkinson, A.B. (1996). The economics of the welfare state. *The American Economist, 40*(2), 5-15.

Barón, J.D., Cobb-Clark, D.A., & Erkal, N. (2015). Welfare receipt and the intergenerational transfer of work-welfare norms. *Southern Economic Journal, 82*(1), 208-234.

Barr, N. (2020). *The economics of the welfare state* (6th ed.). Oxford: Oxford University Press.

Barr, N. (1998). *The economics of the welfare state* (3rd ed.). Oxford: Oxford University Press.

Becker, G.S. (1957). *The Economics of Discrimination* (1st ed.). Chicago: University of Chicago Press.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *The Journal of Economic Perspectives, 30*(3), 141-164.

Dreber, Anna., & Magnus Johannesson. (2019). Statistical significance and the replication crisis in the social sciences. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford: Oxford University Press.

Esping-Andersen, G. (1990). *The three worlds of welfare capitalism.* Princeton: Princeton University Press.

Folbre, N., & Goodin, R.E. (2004). Revealing altruism. *Review of Social Economy, 62*(1), 1-25.

Fong, C.M., Bowles, S., & Gintis, H. (2006). Strong reciprocity and the welfare state. In S. Kolm & J.M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity* (Vol. 2, pp. 1440-1462). Amsterdam: Elsevier B.V.

Freeman, G.P. (1986). Migration and the political economy of the welfare state. *The Annals of the American Academy of Political and Social Science, 485*(1), 51-63.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641-651.

Gilens, M. (1995). Racial attitudes and opposition to welfare. *The Journal of Politics, 57*(4), 994-1014.

Guala, F. (2019). Preferences: Neither behavioural nor mental. *Economics and Philosophy, 35*(3), 383-401.

Harkness, S.K. (2016). Discrimination in lending markets: Status and intersections of gender and race. *Social Psychology Quarterly, 79*(1), 81-93.

Hawkins, S.A., Hoch, S.J., & Meyers-Levy, J. (2001). Low-involvement learning: Repetition and coherence in familiarity and belief. *Journal of Consumer Psychology, 11*(1), 1-11.

Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioural and Brain Sciences, 33*(2-3), 61-83.

Hochman, H.M., & Rodgers, J.D. (1969). Pareto Optimal Redistribution. *The American Economic Review, 59*(4), 542-557.

Holcombe, R.G. (2018). Gordon Tullock on inequality and redistribution. *The Independent Review, 23*(2), 227-247.

Iida, T., & Matsubayashi, T. (2010). Constitutions and public support for welfare policies. *Social Science Quarterly, 91*(1), 42-62.

Jæger, M.M. (2009). United but divided: Welfare regimes and the level and variance in public support for redistribution. *European Sociological Review, 25*(6), 723-737.

Kam, C.D., & Nam, Y. (2008). Reaching out or pulling back: Macroeconomic conditions and public support for social welfare spending. *Political Behaviour, 30*(2), 223-258.

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology, 55*(2), 187-195.

Larsen, C.A. (2008). The institutional logic of welfare attitudes: How welfare regimes influence public support. *Comparative Political Studies, 41*(2), 145-168.

Levitan, S.A. (1985). How the welfare system promotes economic security. *Political Science Quarterly, 100*(3), 447-459.

Lundström, B. (2020). *Study destination: Stockholm. International student mobility in 2018-2019.* Stockholm Academic Forum. https://www.staforum.se/wp-content/uploads/Study-Destination-Stockholm-International-Student-Mobility-2018-2019.pdf

Margalit, Y.M. (2013). Explaining social policy preferences: evidence from the great recession. *The American Political Science Review, 107*(1), 80-103.

Meltzer, A.H., & Richard, S.F. (1981). A rational theory of the size of government. *Journal of Political Economy, 89*(5), 914-927.

Mendelberg, T. (2001). *The race card: campaign strategy, implicit messages, and the norm of equality.* Princeton: Princeton University Press.

Moene, K.O., & Wallerstein, M. (2003). Earnings inequality and welfare spending: a disaggregated analysis. *World Politics, 55*(4), 485-516.

Monroe, K.R. (1996). *The heart of altruism*. Princeton: Princeton University Press.

Offe, C. (1987). Democracy against the welfare state? Structural foundations of neoconservative political opportunities. *Political Theory, 15*(4), 501-537.

Ostrom, E. (2001). Collective action and the evolution of social norms. *Journal of Economic Perspectives, 14*(3), 137-158.

Papadakis, E., & Bean, C. (1993). Popular support for the welfare state: A comparison between institutional regimes. *Journal of Public Policy, 13*(3), 227-254.

Petersen, M.B., Skov, M., Serritzlew, S., & Ramsøy, T. (2013). motivated reasoning and political parties: Evidence for increased processing in the face of party cues. *Political Behaviour, 35*(4), 831-854.

Roberts, K.W.S. (1977). Voting over income tax schedules. *Journal of Public Economics, 8*(3), 329-340.

Romer, T. (1975). Individual welfare, majority voting, and the properties of a linear income tax. *Journal of Public Economics, 4*(2), 163-185.

Roosma, F., Gelissen, J., & van Oorschot, W. (2013). The multidimensionality of welfare state attitudes: A European cross-national study. *Social Indicators Research, 113*(1), 235-255.

Sandmo, A. (1995). Introduction: The welfare economics of the welfare state. *The Scandinavian Journal of Economics, 97*(4), 469-476.

Sihvo, T., & Uusitalo, H. (1995). Economic crises and support for the welfare state in Finland 1975-1993. *Acta Sociologica, 38*(3), 251-262.

Siegel, S., & Castellan, N.J. (1988). *Nonparametric statistics for the behavioural sciences* (2nd ed.) New York: McGraw-Hill.

Soroka, S., & Wlezien, C. (2014). Economic crisis and support for redistribution in the United Kingdom. In Bartels, L, & Bermeo, N. (Eds.), *Mass Politics in Tough Times* (pp. 105-127). Oxford University Press.

Stegmueller, D., Scheepers, P., Rossteuscher, S., & de Jong, E. (2012). Support for redistribution in western Europe: Assessing the role of religion. *European Sociological Review, 28*(4), 482-497.

Sunstein, C.R. (2016). Do people like nudges? *Administrative Law Review, 68*(2), 177-232.

Tullock G. (1967). The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal, 5*(3), 224-232.

Tullock, G. (1971). The charity of the uncharitable. *Economic Inquiry 9*(4), 379-92.

Van Oorschot, W. (2000). Who should get what, and why? On deservingness criteria and the conditionality of solidarity among the public. *Policy & Politics, 28*(1), 33-48.

Young, H.P. (2015). The evolution of social norms. *Annual Review of Economics, 7*, 359-387.

9. Appendix

9.1 Experimental Instructions

This appendix consists of the various communications as well as the survey taken by students in the course of this experiment. Annotations have been added in italics to clarify the structure of the survey. A link to the full survey in PDF form is also available on the OSF preregistration for this experiment (linked in a prior footnote).

The text below is the e-mail received by students at the Stockholm School of Economics asking them to participate in the experiment:

Subject Line: SSE Survey Experiment Participation

Hello!

I'm conducting an experiment for my master's thesis here at SSE and need participants. All that's required is to answer a short Qualtrics survey with one main question. The estimated completion time is **2 minutes** and by participating you have a chance to win one of six lots of **500 SEK**.

Your participation will be anonymous to other participants, and totally anonymous unless you're randomly drawn to win.

The survey can be accessed on desktop or mobile with the following link:

https://hhs.qualtrics.com/jfe/form/SV_bEEVnq3GWaxE3bv

Please keep in mind that participation is meant to be individual, and thank you very much for participating if you do!

Best regards,

Markus Jury

41216@student.hhs.se

A follow-up email was drafted to send out a week before the end of the experimental period in case responses were below the desired sample size, but this proved not to be an issue and thus it was not used.

When the participant clicked on the enclosed link, they were directed to the Qualtrics website, to the survey containing the experiment:

[Page 1]

[**Introduction**]

Welcome to the survey!

This is a short survey contributing to a Master's thesis at SSE. The expected time for completion will be about 2 minutes. Participation is anonymized - your responses will not be linked to your emails. The prize draw for **6 lots of 500 SEK** will be conducted from the email list by a third party. Please do not discuss the experiment with others after your participation until the response period has ended on the 28th of March, 2020. Your participation is very much appreciated!

[Page 2]

[**Online Consent Form**]

Online Survey Consent Form

Purpose of Research: To investigate economic decision-making

What you will do in this research: You'll participate in a decision-making study by answering some survey questions.

Time required: Participation will take approximately 2 minutes to fill out one (1) main survey question.

Risks: There are no known risks associated with this study. Your responses will be anonymized, as will your participation unless you are selected to win a prize.

Benefits: 6 participants will be selected randomly to receive 500 SEK if they complete the survey.

Anonymity: Your responses as well as your participation will be anonymous. The only exception is if you win the prize draw, in which case payout requires that your participation is known to the experimenter.

Participation and withdrawal: Participation is voluntary and you may quit the survey at any time.

To contact the researcher: This study is being conducted by Markus Jury from the Stockholm School of Economics. If you have questions about the research, please contact Markus at 41216@student.hhs.se
By selecting "I consent" below you indicate you have read and understood these conditions and agree to participate in the survey.

- ○ I consent to taking this survey. (1)
- ○ I do not consent. (2)

If the students chose 'I do not consent', they were redirected to the end of the survey page and did not submit their email or any other data.

[Page 3]

[**Email Submission**]

Please enter your SSE student email here for verification & prize-entry purposes.

_____

At this point the students entered the 'Experimental Question Block', and were randomly assigned to one of Pages 4, 5 or 6 as assignment to their experimental treatment. Page 4 corresponds to the 'no cue' condition, Page 5 the 'deserving' condition, and Page 6 the 'lazy' condition.

[Page 4]

[**No Cue Condition Experimental Question]**

Imagine a person who is on social welfare.

Should social welfare provision continue to benefit people like them?

- ○ Strongly Disagree - No, It Should Not (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree - Yes, It Should (7)

[Page 5]

[**'Deserving' Condition Experimental Question]**

Imagine a person who is on some form of social welfare. They have worked regularly in the past, but suffered a work-related injury. They have recovered and are very motivated to get back to work.

Should social welfare provision continue to benefit people like them?

- ○ Strongly Disagree - No, It Should Not (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree - Yes, It Should (7)

[Page 6]

['Lazy' Condition Experimental Question]

Imagine a person who is on some form of social welfare. They are fit and healthy, but have not regularly held a job. They are not motivated to find work.

Should social welfare provision continue to benefit people like them?

- ○ Strongly Disagree - No, It Should Not (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree - It Should (7)

Due to an unfortunate typo the labelling on the 'Lazy' Condition Likert Scale response, it differed slightly from the other two – noting 'It Should' rather than the clearer 'Yes, It

Should' next to the 'Strongly agree (7)' option. This is discussed in the limitations section.

Upon completion of the experimental question, respondents were directed to the end of survey message that read as follows:

[Page 7]

[**End of Survey Message**]

Thank you very much for taking this survey! If you entered your unique SSE email at the start of the survey, you have been entered into the prize draw and will be contacted via that email once the survey period is over (so after the beginning of April) should you win.

If you have any further questions, please don't hesitate to contact me at 41216@student.hhs.se

---

Once the survey period ended on the 28th of March, students opening the survey page were shown the following message instead:

Sorry! The survey period for this survey is over. Thank you for your interest!

If you have any further questions, please don't hesitate to contact me at 41216@student.hhs.se

---

9.2 Pre-registration

This appendix is a copy of the OSF pre-registration analysis plan that was published to the OSF on the 24th of February, 2020, and left embargo after the end of the survey period on the 28th of March, 2020. No changes were made to the experimental design between publishing the pre-registration and running the experiment.

Title: The Influence of Perceived Recipient Deservingness on Welfare Approval

Description:

This study investigates whether perceptions about welfare recipients changes support for welfare programs. Much economic literature has been devoted to the assumption that individuals' evaluation of welfare programs is based on calculated self-interest. The Romer/Meltzer/Richard framework argues that support for redistribution grows as median income falls (Meltzer & Richard, 1981). Extensions such as Moene & Wallerstein's 2003 incorporation of a social insurance

explanation also follows the self-interest route. Following these models, the perceived characteristics of other recipients of welfare should make no difference to individuals' support of such programs.

Aaroe & Petersen (2014), working from a basis in political science, discuss several other explanations for differences in welfare support among different populations, particularly institutional path dependencies (Larsen, 2007). The thrust of their argument, however, focuses on the micro-level influence of psychology on welfare support. Discussing the concept of a 'deservingness heuristic', they emphasize the possibility of a role for "differences in available information", i.e. media messaging and resulting opinion formation, in changing individuals' opinion on welfare (Aaroe & Petersen, 2014, p.686).

Aaroe & Petersen (2014) investigate this matter by conducting a cross-national experiment between Danish and American populations, arguing that despite the dissimilar stereotype predispositions of each nationality once individuals are exposed to 'deservingness cues' these differences melt away.

This study intends to conduct a similar experiment among a single population, students at the Stockholm School of Economics, to see if their core observation is replicable in that environment.

The experiment consist of one control group, which receives no information as to a hypothetical recipient's deservingness, and two treatment groups. The first treatment group receives cues that depict a hypothetical recipient as 'unlucky', while the second receives cues that depict a hypothetical recipient as 'lazy'. Whether or not these results produce significant differences in individuals' approval of welfare programs helping these hypothetical recipients would indicate whether individuals' perceptions of welfare recipients affect their approval for welfare programs.

Hypotheses:

The primary hypothesis of this paper is that there will be a significant difference in support for welfare programs between experimental groups.

I expect that the first treatment group ('unlucky' recipient cues) will show the highest average support for such programs, the second treatment group ('lazy' recipient cues) the lowest, and the 'no information' group somewhere between the two. Pair-wise, this means I expect the mean result from the first treatment group to exceed that of the control group and the second experimental group, and the mean of the control group to exceed that of the second experimental group.

The expected effect size is difficult to quantify, as Aaroe & Petersen (2014) perform their analysis with a regression on the respondents' country of origin and do not report a group or pairwise analysis of the basic experimental conditions' (deservingness cues) effects. Considering the extensive literature supporting deservingness heuristics playing a large role in determining public opinion of welfare (Fong, Bowles & Gintis, 2006, pp. 1439-1461), I expect the true effect size may still be large enough to provide useful results with my sample size.

Design Plan

Study Type:

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Blinding:

For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

Study Design:

On a more specific level than in the introduction, I am seeking to evaluate whether simple deservingness cues (or lack thereof) about a hypothetical welfare recipient changes the average response to a question about whether social welfare benefits should continue to go to recipients like the hypothetical person.

Much of this experiment is focused on replicating and testing a portion of Aaroe & Petersen's 2014 paper "Crowding Out Culture: Scandinavians and Americans Agree on Social Welfare in the Face of Deservingness Cues", published in The Journal of Politics. Aaroe & Petersen's paper focused on cross-cultural differences in stereotypes and the ability of cues to counteract the effects of these stereotypes; this paper differs in that it seeks to replicate the more fundamental question of whether such cues significantly influence respondents' approval of social welfare benefits going to the person.

This paper further differs from Aaroe & Petersen in methodology: while the two researchers used OLS regressions across national groups, this paper intends to use the Kruskal-Wallis H Test followed by the Mann-Whitney U for pairwise comparisons, should the Kruskal-Wallis find significant or indicative results.

The experiment will be conducted via an online Qualtrics survey, with recruitment occurring through an email to a list of approximately 1700 students current at the Stockholm School of Economics. Prior to answering the experimental question, students will be asked to confirm a standard consent form, be informed of the possibility of winning a 500 SEK prize for participation, and be asked to report their SSE email in the survey itself for verification and prize-drawing.

The experiment itself will contain three treatments:

- Treatment 0 (the control)

- Treatment 1 (the 'unlucky' recipient cue condition)

- Treatment 2 (the 'lazy' recipient cue condition)

Treatment 0, the control group, will be asked a question about approval of social welfare for an individual about whom they are given no deservingness cues, as follows:

"Imagine a person who is on some form of social welfare."

Their response will be selected from a multiple-choice Likert scale, ranging from a value of 1 for Strongly Disagree (that people like the individual should continue to benefit from social welfare) to 7 for Strongly Agree.

Treatment 1, the 'unlucky' group, will be given the following description:

“Imagine a person who is on some form of social welfare. They have worked regularly in the past, but suffered a work-related injury. They have recovered and are very motivated to get back to work.”

The response choices will be the same as for question 1.

Treatment 2, the 'lazy' group, will be given the following description:

“Imagine a person who is on some form of social welfare. They are fit and healthy, but have not regularly held a job. They are not motivated to find work.”

The response choices will once again be the same.

Students will be randomly assigned one of these three questions through Qualtrics' question randomization option. The survey will be sent out in English and only English on the assumption that students at SSE will have a fluent command of English; SSE's Master's programs are conducted entirely in English and from the fall of 2020 so will its Bachelor's programs.

The survey will run from the 24th of February to the 28th of March, a period that totals 4 weeks and 5 days. The e-mail to students will inform them of the payout structure (random draw of 6 email addresses, each receiving 500 SEK unconditionally), and it as well as the survey introduction attempt to dissuade discussion of the survey until it has run its course. The e-mail, survey and survey messages are attached.

Randomization:

The survey tool used for this experiment, Qualtrics, has an option that allows for randomization of which questions are presented to an experimental respondent. A respondent will be assigned and only see one of the three treatment questions once they have opened the survey and read & agreed to the instructions.

Sampling Plan

Existing Data: Registration prior to the creation of data

Explanation of existing data: N/A

Data collection procedures:

The data for this experiment will be collected through a Qualtrics survey sent out to students' university e-mail accounts at the Stockholm School of Economics. Respondents will be excluded if

they do not consent to the experiment on the initial survey landing page, if they do not answer the main survey question, or if they fail to provide a unique SSE student email at the start of the survey.

The email containing a link to the survey as well as the survey landing page will give a brief overview of the task and inform them that if they complete the survey they will be entered into a random drawing to win 6 allotments of 500 SEK; assuming 300 participants this works out to an expected payment value of 10 SEK per participant. Payment will take place via bank transfer once the experiment has concluded. Eligibility for payment will follow the exclusion criteria above.

The survey will be emailed out to a list of approximately 1700 students on the 24th of February, and run until the 28th of March. During this time the number of respondents will be monitored through Qualtrics to estimate the eventual sample size, but data on responses to the questions will not be visible to the experimenter.

If at the start of the last week of the survey (the 21st of March) the number of respondents has not yet reached the target sample size of 300, the experimenter will send out a reminder email.

Sample Size:

The target sample size for this study is at least 300 respondents. Prior studies using a similar methodology suggests this is an attainable sample size. Due to the randomization method used, this should result in approximately 100 respondents per experimental group. If the response rate is high enough that the sample size is larger this will only increase the strength of the analysis, and will be welcomed.

Sample Size Rationale:

With the survey sent out to approximately 1750 students, a target sample size of 300 assumes a response rate of approximately 17%; this response rate has proved a reliable target in survey experiments conducted in a similar manner over the past year at SSE.

The target sample size has been set at 300 in order to ideally gather sufficient experimental power to prove or disprove the experimental hypothesis, as otherwise the results would be susceptible to type M or S errors (Gelman & Carlin, 2014).

Stopping Rule:

The termination of data collection will occur at the pre-specified ending date of the survey, regardless of how many responses have been collected.

Variables

Manipulated Variables:

Measured Variables:

The measured variable will be respondents' answers to the survey question, i.e. their support for social welfare in the presence (or absence) of deservingness cues. Operationally, they will choose one of seven options ranging from Strongly Disagree through to Strongly Agree.

Indices:

The stages of response to the survey question are scaled by a Likert scale as follows:

Strongly Disagree - 1

Disagree - 2

Somewhat Disagree - 3

Neither agree nor Disagree - 4

Somewhat Agree - 5

Agree - 6

Strongly Agree - 7

In order to find means for each group, their answers will be summed via Likert scale value and then divided by the n of each group. This will allow a comparison of the mean of each group in both group and pairwise comparisons.

Analysis Plan

Statistical Models:

The hypothesis of this experiment is that there will be a significant difference in the response means for the different experimental groups, i.e. groups that are given different deservingness cues about social welfare will display different levels of approval for social welfare.

The three experimental groups at the Control Group (0), which is given no cues; the 'Unlucky' Group (1), which is given a cue that signals deservingness; and the 'Lazy' Group (2), which is given a cue that signals undeservingness.

Once the means of each group's response has been collected (see Indices section under Variables), a Kruskal-Wallis H Test will be run across all 3 treatment groups. The Kruskal-Wallis H test is a non-parametric extension of the Mann-Whitney U test to more than two groups. Its null hypothesis is that k sampled populations have the same average or median; the alternative hypothesis being that at least one sample is from a distribution with a different average/median. As it is a nonparametric test, it does not require an assumption of a normal distribution in responses from the population, which is something that can’t be assumed with a behavioural response question.

If the Kruskal-Wallis H Test returns a significant ($p<0.005$) result in rejecting the null hypothesis, each groups' average will be compared with the other groups using pairwise Mann-Whitney U tests. Should the test return an indicative but not a significant result ($p<0.05$), the Mann-Whitney U tests will proceed with their results considered as suggestive evidence.

Transformations:

No transformations of the data should be required beyond the compilation of group averages.

Inference criteria:

As discussed above, the results of tests will be treated as significant if they are under the p-value threshold of 0.005. If the results show this, it means that the average response to questions differed between groups' different questions. They will be treated as suggestive evidence if they are under the more traditional p-value threshold of 0.05. Should the results not reach either value, it means that there was no difference in average response to the groups' different questions.

The Kruskal-Wallis H Test & Mann-Whitney U pairwise tests being conducted are two-tailed.

Data exclusion:

Data points will be excluded on a number of criteria:

To prevent duplicate submissions or participants who are not in the target population, responses will only be included if they also include a valid, unique SSE student email. Should the same email submit multiple responses, all their responses will be excluded, as it is likely they have not seriously answered the question at that juncture.

Missing data:

As the survey requires a response to the experimental question, there is no possibility of 'missing' data.

Exploratory Analysis:

There is no exploratory analysis planned for this paper.