CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Normative expectations and subjective beliefs: an incentivised experimental study
Authors	Wang, Cuizhu
Publication date	2022-12-12
Original Citation	Wang, C. 2022. Normative expectations and subjective beliefs: an incentivised experimental study. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2022, Cuizhu Wang. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2023-07-23 17:59:43
Item downloaded from	https://hdl.handle.net/10468/14502

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Normative Expectations and Subjective Beliefs: An Incentivised Experimental Study

Thesis presented by

Cuizhu Wang, M.A., B.A., and B.B.A.

0000-0002-4703-3323

for the degree of

Doctor of Philosophy

University College Cork

School of Society, Politics & Ethics

Department of Philosophy

Head of School & Department: Professor Don Ross

Supervisors: Professor Don Ross & Professor Glenn Harrison

2022

Table of Contents

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Abstract

This thesis is an experimental study to investigate the operationalisability of the theory of social norms provided by Cristina Bicchieri.

In Chapter 1 I critically summarise a main theme from recent literature and distinguish the accounts of norms based on social preferences from accounts based on social structure. I also summarise different theorists' accounts of social norms as a social construct, in addition to surveying some issues scholars have raised empirically.

Chapter 2 reviews the conceptual analysis of social norms by Bicchieri as a social structure based account. Bicchieri's conceptual analysis introduces three kinds of condition for norm identification. I review these in detail, and suggest hypothesis testing corresponding to each kind of condition. Chapter 2 briefly analyses a critical problem for Bicchieri's theory.

Chapter 3 provides philosophical background that supports the intentional concepts applied in Bicchieri's analysis of social norms. I suggest that the Dennettian account of Intentional Stance is the best philosophical framework for Bicchieri's account of social norms. I also argue that Revealed Preference Theory from economics is an application of the Intentional Stance. I conjecture that adopting the Intentional Stance and applying Revealed Preference Theory to empirical data can allow for improved operationalisation of Bicchieri's conceptual analysis.

In Chapter 4 I provide critical review of some key experimental work by Bicchieri and co-authors applying her conceptual analysis of social norms. I then provide a critical review of a widely used toolbox from the current economic literature for norm elicitation. Then I introduce a more rigorous experimental protocol for investigating social norms understood following Bicchieri's analysis. The toolbox suggested in my thesis addresses limitations identified in Bicchieri's empirical work.

Chapter 5 presents design of the experiment administered as the core element

of the thesis.

Chapter 6 shows and analyses the results from the experiments described in Chapter 5. It also introduces the statistical models used in my thesis to assess the extent to which Bicchieri's analyses successfully guides experimental identification of social norms.

Chapter 7 offers concluding theoretical reflections, and discusses possible extensions of the research presented in this thesis.

Acknowledgements

My greatest debt is to my Ph.D. supervisors Professor Don Ross and Professor Glenn Harrison. Over the years of learning from them as a Ph.D. student, I consider myself a very lucky "academic daughter" who had the luxury to receive their strict training, their earnest instructions, and their invaluable, insightful, and professional advice. Their immense knowledge, their passionate pursuit of good scholarship, and their insistence on being responsible as knowledge producers through extremely disciplined everyday research and "dreary" re-evaluations and re-corrections has exemplified for me what a good scholar is, and how much it takes to be one.

The deepest kindness I received from my supervisors not only comes from their persistence in teaching me, but also their generosity in pulling together resources to support my thesis project. It was only made possible with the funding they arranged from the Center for the Economic Analysis of Risk (CEAR) at Georgia State University and the collaboration of the University of Cape Town.

I am thankful of colleagues from CEAR who have provided support, and I am also extremely grateful to Professor Andre Hofmeyr, Professor Karlijn Morsink, and Dr. Brian Monroe who provided enormous and essential help on the experimental project for my thesis. My extended thanks also go to Dr. Nelleke Bak who provided help throughout the years behind Professor Ross, and Professor Elisabet Rutström who put up with Professor Harrison's commitment at the end of my Ph.D. by accepting his trip to Cork at a difficult time in May 2022. My further thanks go to Professor Tadeusz Zawidzki for his mentoring through Zoom calls on his theory of Mindshaping.

I would like to also express my deepest gratitude to the International Office, College of Arts, Celtic Studies & Social Sciences, and Department of Philosophy at University College Cork (UCC). Different offices at UCC have together provided the financial support to my Ph.D. project. Specifically, I would like to thank Dr. Alan Gibbs, Miss Anne-Marie Scarry, and Mr. Conor Delaney for institutional support.

The Department of Philosophy have provided stimulating scholarly support, of which I am particularly thankful to Dr. Ouyang Xiao, a previous post-doc at the Department, who guided me to this path of pursuing a Ph.D. in Philosophy. I am also thankful for senior colleagues at the Department for their professionalism and their constant support.

I sincerely and forever thank my family, Fengzhu Wang, Longzhu Wang, and Yiming Ma. It is their forever faith in and love for me which gives and constantly reminds me of the primary courage for pursuing this path. In this family team, I am also deeply indebted to my childhood friend Li Dan, and Professor Dr. David Johnson from Ferrum College, Virginia, USA, both of whom have always loved me and looked out for me since the start of our friendships.

Last but not least, I would like to also thank my good friend Professor Mohamed Siala from INSA Toulouse and LAAS-CNRS for being such an inspiration and a great scholarly influencer to me since 2018. I also would like to thank my close friends Dr. Marten Kaas, Sarah Romer, Giacomo Melloni, Kornela Andruszkow, Dr. James Cronin, and Carolyn O'Brien for the invaluable friendships. I further thank their support in different aspects at the end of my Ph.D.: the hardships would never be possibly conquered without these friends while I am being blocked away from my family in China due to COVID restrictions.

Use of Personal Pronouns

In this thesis I adopt the convention of using "she" for third-person singular pronouns unless the gender of an actual person is obvious from the context, or pronouns occur within direct quotations of other authors.

Introduction - Norm-shaped Minds and Behaviours

Imagine that I am hosting a Chinese hotpot dinner for friends from Canada, Germany, Tunisia, Poland, Italy, Ireland, and USA. We were all born at different times, and we grew up in different family and cultural backgrounds before we met. We now are dedicated to different life goals and we hold different roles in society. We think and act differently. In a situation like this, how should each of us behave? What topics should we bring to the conversation to engage others? What kinds of topics should we avoid? What way of interacting would be taken as offensive by some members at the table who come from different corners of the world? It is a typical scenario under globalisation in which many social norms can be at play at the same time as we interact with others.

Social norms create what Bicchieri (2006) calls a "heuristic route". According to the heuristic route, norm compliance is an automatic response to situational cues that focus our attention on a particular norm, rather than a conscious response that gives priority to normative considerations. Social norms dictate our behaviours so automatically that we might be unaware of them most of the time. An example given by Bicchieri (2017) is that a father who sends his daughter to school may not recognise that this will likely have implications for perceptions of her honor and purity in his traditional culture.

A metaphor I usually give to people about my attempt to address the question of what a social norm is is to suggest people to imagine a social norm as a virtual wall. When we enter a city, such as York in the UK or Xi'an in China, we see a wall surrounding the ancient city center. The old city walls remind us that if we had entered the city when the city walls still played their intended institutional roles, we would have been expected to behave differently inside than outside. Expectations have a similar motivational force in the imagined virtual walls that demarcate social norms.

When we travel to a new country with a different cultural heritage from our

own, we seem to automatically seek for confirmation of the behaviours are expected and/or considered to be appropriate in certain situations. Such confirmations typically result from inferences we make from observing how others behave and respond to behaviours they observe. Circumstances like this, which requires "norm-hunting effort", are so common that we can find them every day within our own cultures without having to look beyond political or cultural borders. In our daily lives, they can simply be found when we shift social circles.

Awareness of the existence of social norms may only be made explicit in situations when the norms are foreign to us. When we stay within "the virtual wall", we hardly see how many behaviours and beliefs are products of the fact that we are in it. As long as our social interactions are going smoothly, self-reflection is usually unnecessary. However, we may not realise that the governance by the norms that we follow in our daily lives can cause us troubles. Imagine receiving a foreign guest at your house. You may be so open-minded and considerate as to research the kind of breakfast that guest would be comfortable with, if you know that your guest comes from a culture where people do not eat the typical breakfast consumed in your household. If the guest leaves unfinished breakfast you had put so much effort into preparing, you may interpret your guest's behaviour as indicating dislike of your food service, or you might consider her behaviour as being thoughtlessly rude. But perhaps, your guest actually intentionally left some unfinished morsels on the plate because in her culture it is regarded as rude to completely finish a dish as a guest. Perhaps she acted on the belief, carried over from her cultural history, that she was meeting your expectations as a host.

Such misunderstanding of expectations due to social norms can also arise among people from a common culture. Imagine that you travel to another country where most people do not speak your language. Then imagine that one day, you happen to meet a person who speaks your language, and that you even know about her hometown well. You are likely to feel close to her immediately. You enjoy the easiness of not needing to "explain yourself". However, there is a problem with this

easiness. You assume you understand each others' nonverbal behaviour, as you both are familiar with the behavioural cues which are generally shared across your home country. Unfortunately, as your interactions deepen over time, this compatriot seems to constantly fail to act upon your expectations. You may feel disappointed and start to negatively judge the character of this person. However, the truth is that as this person has adjusted to life abroad, she has developed new behavioural cues. The easiness of "not explaining ourselves" starts to fall apart. Your feelings might reverse their initial polarity. It seems the easiness resulting from the presumed common ground between you and your compatriot has become a burden, a source of misunderstanding. In fact, expectations in this scenario form the underlying barrier that blocked you and your compatriot to make efforts to get to know each other as individual human beings. Instead, you were both blinded by your expectations, the supposed common ground here.

Under the current stage of globalisation in which many of us live, we see growing tendencies to discrimination and polarisation across different social groups and cultures. Perhaps, social norms have contributed to this problem due to the unintentional and unconscious enforcement of expectations each individual aims to comply with, to different degrees. Globalisation has allowed people to *see* and *communicate* more, from which we can reach a deeper understanding of cultural differences. However, what globalisation has not done enough is to enlighten us to see beyond culturally shaped *behaviour*. We must also gain awareness of the process of how our *minds* are shaped differently across different cultures and social circles. We must update our expectations given such understanding. It is the unconsciously and constantly re-shaped minds we must try harder to discover.

Social norms are double-edged swords. On one side, social norms are institutions in which individuals' actions are guided towards cooperation. Social norms hence glue a society together and enable it to maintain its structure. On the flip side, normative forces generated by social norms may harm us at both the individual and societal levels. In this sense, social norms can diminish human

welfare, and the stronger the normative force a social norm carries, the more severely it risks interfering with individual flourishing.

This Ph.D. thesis is an attempt to contribute to understanding what social norms are and how we can model them through scientific investigation. It is my ambition that this learning journey will enhance understanding of human society. I hope that this understanding will lead to a potential contribution to better shared lives.

Chapter 1 - General Background

1.1 Social Preferences versus Social Structure models of Social Norms

In economics, social norms are modelled either as social preferences or social structures. Social preference accounts hold that a person's social conduct is motivated not only by her own payoffs but also by her views about which arrangements of social resources are best. Social structure accounts of social norms, on the other hand, model norms as social facts that agents encounter as elements of their social environments. Such accounts view social norms as facts at the social level which do not reduce to the preferences of individuals. Social norms are regarded as exogenous to agents.

Social preference theory assumes that norms, as social properties, are reducible to the psychological properties of individuals. Social preference accounts (e.g., Fehr and Schmidt 1999; Fehr and Gächter 2000a, 2000b; Henrich et al. 2005) model norms as strictly emergent from individuals' preferences over distributional social properties such as equality, equity, or reciprocity. For example, Fehr and Schmidt (1999) model the norm of equality as arising from "inequality aversion" in individuals. This model suggests that some individuals follow the norm of equality because they suffer from utility loss when they encounter unequal payoff distributions in a social setting. Gintis (2009) argues that social preferences in humans in general are what facilitate cooperation and exchange.

Extensive empirical studies have been conducted to examine models of pro-sociality following the social preference hypothesis. For example, Dana, Weber and Kuang (2007) (DWK) conducted an experiment using a Dictator Game to understand the determinants of fair behaviour. They introduced four treatments in which subjects were left with some moral "wiggle room" to behave self-interestedly. The main difference between the four treatments in their study is that in their baseline treatment there was full transparency between subjects' actions and outcomes, whereas the other three treatments manipulated different degrees of elimination of

the transparency. The main idea behind the experimental design is that if giving in the transparent baseline game reflects a preference for an equal split among the players, then in hidden payoff settings, the proportions of what Dictators give should be equal to the amounts when the decisions of the splits are transparent. Instead, by having wiggle room (allowed by the treatments except the baseline) and excuses to not feel compelled to give an equal amount, subjects take advantage of the chances to behave under ignorance in order to maximise their own self-interested payoffs. The results turned out to favor their prediction that a moral wiggle room undermining agents' compliance to the equality norm. This contradicts the social preference hypothesis that people have social preferences (about inequality aversion) in their sense. Dana, Cain, and Dawes (2006) and Lazear, Malmendier, and Weber (2012) found similar results.

Social preferences as internalised psychological features in individuals no doubt exist. However, modelling social norms as social preference accounts falls into the pitfall which Gilbert (1989) calls "the simple summative account" of social facts. It ignores the scale relativity of ontology.[1] This feature may be explained by the description by Ross (2022) that social preference accounts of social norms "encourage a psychological style of experimental design, in which researchers set participants in games featuring equilibria that vary in these distributional properties and seek to identify individuals' characteristics that predict preferences over them" (p. 19).

Binmore (2005, 2006, 2010) points out two major problems with the social preference approach at its operational level, which again threaten its generality. First, the social preference model allows for too much freedom of specification in utility functions. As the utility model extends to situations where multiple norms are potentially at play, the model becomes empirically empty. Imagine a situation where

[1] Scale relativity of ontology is the idea that which terms of description and principles of individuation we use to track the world vary with the scale at which the world is measured (Ladyman and Ross 2007, p. 199).

norms of trust, gender equality, reciprocity, and race equality intertwine with each other. Then we may not be able to judge which norm is motivating an agent's behaviour. If we follow the social preference model, we will have to write the utility model to include all the potential norms at stake, as a deduction from the agent's payoff argument. In an age of globalisation where in increasing ranges of contexts we hardly are sure how many norms can be at play in a social scenario, it can turn out the utility functions suggested by social preferences models are impossible to identify.

Another problem is that in the typical experimental design subjects are forced to behave according to the norm assumed to be at stake. For example, in Fehr and Schmidt (1999), the experiment was designed such that only one norm (norm of fairness) is potentially relevant to the subjects in the task. This induces the experimental result to reveal the social preference at stake. Another example is that in Fehr and Rockenbach (2003), the experiment was designed to allow subjects to impose punishments for violations of a norm of cooperation. Consequently, subjects' utility models written in these experiments are interpretated as a form of confirmation of the norm of cooperation. This way of designing experiments assumes that multiple norms are not at play in subjects' utility models by ruling them out in the way the experiments are designed. But this assumption lacks general validity.

Social structure accounts, on the other hand, model social norms as social facts (Gilbert 1989), which are irreducible to individuals' psychology. Georg Simmel (1895) raised the formulation of the concept of social structure as a web of "crystallized relationship". Durkheim (1982) sees social phenomena as special entities, distinctive from the actions of individuals which constitutes their social interactions with each other. Durkheim names social phenomena as "social facts". According to Martin (2009), followers of Simmel and Durkheim on social structures hold the following general view:

> "Social interactions, when repeated, display formal characteristics; and this form can then take on a life of its own, ultimately leading to institutions that we (as actors) can treat as given and exogenous to social action for our own

> purposes, though at any moment (or at least at some moments) these institutions may crumble to the ground if not rejuvenated with compatible action" (p. 3).

These reflections are not intended as comments on the viability of any general ontological individualism. There cannot be an operative social norm if there are no individuals whose behaviour supports its maintenance. An ontological individualism with bite would go beyond this truism, by maintaining that all social-structural properties, so by implication all such characteristics of norms, can be fully explicated, at least in principle, by reference to properties of, and relations between, individuals. Philosophers typically understand ontological individualism by reference to some analysis of *supervenience*.[2] Supervenience can be interpreted metaphysically, as specified in terms of possible worlds (for example, as in Epstein 2015), or as a contingent relationship between types at different levels that is established provisionally, based on a sample of empirical observations (Guala 2022). Rejecting the supervenience of norms on individual preferences would be a possible basis for rejecting the adequacy of social-preferences models of norms. But that would be a stronger position than is needed for present purposes. Interest in social structure models, as opposed to social preferences models, can be motivated by methodological considerations alone. Furthermore, such departure from *methodological* individualism need not depend on a claim that methodological individualism is a *generally* bad programme for social scientists.[3] I criticise the methodological individualism of social-preferences accounts of norms based on problems that are specific to the goal of empirically identifying and estimating norms from observed choice behaviour; no stand is taken here on any more general philosophical individualism.

Another concept which is entangled with the social structure account of social norms is institution. Martin (2009) points out two features of social structure which

---

[2] See Savellos and Yalçin (1995) for a representative sample of such analyses.

[3] There are such general arguments in the literature. Epstein (2015) makes such a case on metaphysical grounds. Ross (2014) makes such an argument, at least for economic approaches such as the one applied in this thesis, based on the contingent but general objectives and modelling approaches of economists. Zahle and Kincaid (2019) argue that there is no general default in favour of methodological individualism as furnishing superior explanations even when it is possible to adopt it.

leads to an understanding of the notion of institution. First, it seems that there are *conditions* under which the interactions between individuals tend to align with each other. Second, social structure appears to us as patterns of behaviour across a population, at some point the structure starts to take on exogenous reality: it persists and decays, and it regulates human interactions. Theorists in this tradition suggests at this point we must speak of an institution. Martin defines institution as follows, "an institution exists when interactants subjectively understand the formal pattern in terms of the content of relationships" (p. 4).

Institutions are of core interest to some economists, political scientists, and social philosophers. For example, social ontologists (Guala 2016) investigate institutions in general and develops general theories of the abstract type. Economists such as Ostrom (1980) and North (1991, 1993) are leading examples of contemporary institutional economists. North (1991) defined institutions as "humanly devised constraints that structure political, economic and social interaction" (p. 97). Crawford and Ostrom (1995) view institutions as "enduring regularities of human action in situations structured by rules, norms, and shared strategies, as well as by the physical world. The rules, norms, and shared strategies are constituted and reconstituted by human interaction in frequently occurring or repetitive situations" (p. 582).

The concept of institution following this account sees social norms in this sense are institutions. Precisely, it is "a feature of a social structure, that is, an element of prevailing patterns in relationships among recurrently interacting people that constrains and motivates the behaviour of at least some of them, which arises, persists for a finite time, and ends by decay or catastrophic collapse" (Ross, Stirling and Tummolini 2021).

If social norms regulate human behaviours, an operational definition which allows inferences of the existence of social norms being made through modelling observable behaviours is much needed. Bicchieri (2006, 2017) develops a philosophical conceptualisation of social norm which is intended to serve this

function. The definition of social norms by Bicchieri (2017) is as follows:

> "A social norm is a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (empirical expectation), and (b) that most people in their reference network believe they ought to conform to it (normative expectation)" (p. 35).

Bicchieri's conception of social norms is not only an operational one, but most importantly gives a way of avoiding the shortcomings of social preference accounts. First, a norm-based model should be able to directly test how *contexts* which matter for the norm can be studied in the lab and in the field without losing generality over different types of social norms. For example, in similar contexts some behaviours are regulated by the norm of fairness, whereas some behaviours are regulated by a norm of equity. How can we test the causal effects of contexts on norm compliance without falling into a regime of developing multiple models which includes multiple models, i.e., a model of a fairness norm and a model of an equality norm, for example? In fact, evidence suggests that even simple changes to choice contexts are often sufficient to cause people to switch their choices (Engelmann and Strobel 2004). A common practice in the literature is to manipulate the choice contexts and then observe altered behaviour. For example, in Bicchieri and Chavez (2010, 2013) they manipulated expectation due to the importance of expectation in the theoretical framework.

Second, a norm-based model should also be able to show how a non-reducible account of social facts at the group-level can be studied empirically through observable choices of individuals at the individual-level. Bicchieri's theory of social norms, given its emphasis on *expectations*, suggests a methodology for the identification of social norms as social facts from observations of the choices of individuals. According to Bicchieri (2017, p. 71), the alignment between agents' different types of expectations is a necessary but not sufficient condition for norm existence. If this is true, then applications of statistical modelling on beliefs at both group-level and individual-level should give us a way out for identifying norms from observable behaviours.

Social norms are social facts, whereas social preferences are psychological facts. Social norms are better modelled following the social structure account, instead of social preferences account. Social preferences and social norms aren't the same.

1.2 Social Structure model of Social Norms

Different models following the social structure accounts of social norms are developed in the economic literature. The norm sensitivity account of Kimbrough and Vostroknutov (2016, 2018, 2020) (KV), for example, argues that observed behaviour may be consistent with norm dependent preferences, which refers to a preference per se to obey a social norm.

The concept of norm sensitivity in KV's work is previously captured in the utility model of Bicchieri (2006). Bicchieri (2006, p. 52) defines norm sensitivity as the degree to which an individual "embodies one's personal reasons for adhering to the norm". According to Bicchieri (2017, p. 75), an individual's sensitivity to a norm is "inversely related to the relative importance of one's social expectations in motivating compliance", and that it is a type of "motivational power of social expectations". In this sense, norm sensitivity is a necessary condition for norm adherence.

One example arises in the setting of normative intervention against open defecation reported in Bicchieri (2017, p. 85). Though a village has already invested in building toilets, villagers still avoid using them, unless their actions are monitored, and open defecation is sanctioned. Another example is the case of sanitation practices during early phases of COVID-19 reported by Ashraf et al. (2020). Their report reveals that among 2044 respondents, though 60% had access to a private toilet and 11% to a public or community toilet, 92% of the respondents did not change their defecation behaviours in the first 2 months of COVID-19 lockdown. A norm sensitivity account might be applied here, as one way to model and explain the phenomena: in both cases it might be that the villagers in the first case and the

respondents in the second case have low norm sensitivity, which contributed to a low preference to obey the social norm of toilet using. Lower norm sensitivity, as argued by Bicchieri (2017), reduces the *motivational power* of social expectations to cause norm compliant behaviours. Norm sensitivity can therefore be viewed as necessary condition for norm compliance.

KV draw a clear distinction between individuals' norm sensitivity *across contexts* versus *within a context*. *Context* here refers to the range of norms that might be relevant to an agent's behaviour. In a typical game setting where multiple norms might be relevant, norm sensitivity captures the idea that observed heterogeneity might be explained by the fact that individuals have different degrees of sensitivity to *different* norms. For instance, a person may be sensitive to norms of reciprocity, but less so to equality considerations. We then say that her norm sensitivity varies *across contexts*. Norm sensitivity *within a context*, refers to the idea that observed heterogeneity might be explained by the fact that the degree of norm sensitivity towards the *same* norm differs across different individuals. KV (2016) models the latter.

KV hypothesise that sociality is driven not by preferences over payoffs of others, but rather preferences for following well-established social rules. KV set out to develop a model in which they can illustrate how the notion of a norm-dependent utility can account for heterogeneity in prosocial behaviour across a series of standard experiment paradigms, including the public goods game, trust game, dictator game and the ultimatum game.

KV define norm-dependent utility over both own payoffs and own norm adherence. In the context of a 2-player, one-shot dictator, ultimatum, trust, or public goods game, the utility model is as follows:

$$U_i(a_i, a_{-i}) = u_i(a_i, a_{-i}) - \phi_i g(\|u_i, i(a_i) - i(\eta_i)\|),$$

where η denotes a norm, η_i represents a strategy profile (a set of choices) that are prescribed by η as appropriate, for individual i; a_i represents a single choice of action by player i, and a_{-i} represents a single choice by player 2. The final payoff of

player i is $U_i(a_i, a_{-i})$ which is composed from $u_i, i(a_i)$ and $u_i, -i(a_i, a_{-i})$. The first argument $u_i, i(a_i)$ is the part of the payoff of player i obtained from choosing action a_i in the strategy set which is prescribed by η; and the second argument $u_i, -i(a_i, a_{-i})$ is the part of the payoff that results from player i's partner player choosing for player i. For example, in a one-shot ultimatum game task, the utility for the Proposer i would be $u_{i,i(x)} = x$. If a proposal is rejected by the Responder, then the utility for Proposer i would be $u_{i,i(x,R)} = -x$; if a proposal is accepted by the Responder then the utility for Proposer i would be $u_{i,i(x,A)} = 0$. Furthermore, g represents the disunity of deviating from the norm, where g: [0,1] → [0, 1] is a strictly convex increasing function with $g(0) = 0$ and $g(1) = 1$. The parameter ϕ_i indicates the sensitivity of player i to deviations from the norm. If $\phi = 0$, the agent is not at all sensitive to social norms, and maximises her utility by maximising material payoffs. If $\phi \to \infty$, the agent will always follow the norm, no matter what consequences this has for her material payoffs. The absolute value of the payoff difference calculated from the normalisation function || . || refers to disutility of deviation from the norm relative to the player's ideal payoff given her norm sensitivity.

There are two limitations to the norm sensitivity model at the theoretical level. First, it leaves open the possibility that an individual's degree of norm sensitivity towards the same norm (*within a context*, in KV's language) might differ across different circumstances. The second problem is pointed out by Ross, Stirling and Tummolini (2021, p. 2) that this framework "only accommodates norms for which social welfare increases in the number of followers; nor does it allow for an agent to persistently follow a norm she would better off abandoning". For example, in cases of preference falsification as discussed and modelled by Kuran (1995), agents follow norms they would prefer not to because they have false beliefs about others' preferences.

1.3 Summary

This chapter introduced the general background of the studies on social norms. The key point which sets the stage for my thesis is that social preferences and social norms are different things. I also demonstrated a general account of modelling social norms by KV as an example of the social structure account of modelling social norms. As my thesis proceeds, I rationalise the philosophical account of social norms given by Bicchieri (2006, 2017), which proposes a sensible qualitative analysis for modelling the richness of context more scientifically. My thesis is a project to investigate the operationalisability of Bicchieri's philosophical account of social norms. In the next chapter, I review this account.

Chapter 2 - Bicchieri's Theory of Conditional Norms

Bicchieri analyses social norms as social facts and social structures, instead of as aggregations of individual attitudes to social states as in social preference accounts. Following the game-theoretic tradition, a social norm is defined broadly as a Nash equilibrium which is a combination of strategies. As a social norm is defined as an equilibrium, it therefore is "supported by self-fulfilling expectations in the sense that players' beliefs are consistent, and thus the actions that follow from players' beliefs will validate those very beliefs" (Bicchieri 2017, p. 12). Bicchieri (2006, 2017) analyses a social norm as a "behavioral regularity emerging in a mixed-motive game".

> "A social norm is a rule of behaviour such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (empirical expectation), and (b) that most people in their reference network believe they ought to conform to it (normative expectation)" (Bicchieri 2017, p. 35).

The central role of mutually consistent expectations in supporting a norm is a key feature that differentiates this account from the social preferences account of social norms. First, it gives primary emphasis to the co-existence of empirical and normative expectations as the basis for norm compliant behaviour. According to Bicchieri, to confirm the existence of a norm empirical expectations and normative expectations must have joint effects on people's behaviour. In order for a person to play her equilibrium strategy in an interaction, she has to predict what the others' strategies would be. If she is confident in her expectations about the other players' beliefs, and these beliefs are consistent, she can expect the other players' strategies will follow from their beliefs or validate those very beliefs and this simultaneously extends to the other players whose behaviour supports the norm. So social norms are functions of networks of expectations. Meanwhile, the account bridges ascriptions at the social level and individual level and allows social norms to be identified in choice data of interacting individuals.

The analysis of conditionality in Bicchieri's account suggests an operational

way to identify norms from observed behaviour. Specifically, it is the conditional preference. According to Bicchieri, conditionality of preferences implies that an agent may follow a norm in the presence of the relevant expectations (empirical and normative expectations) but disregard it in their absence. For example, Bicchieri and Zhang (2012) investigated the Fehr-Schmidt (1999) model[4] and pointed out that one of the weaknesses of the social preferences account is that it fails to explain individuals' inconsistencies in behaviour across different situations. Bicchieri and Zhang (2012) argue that this is because the social preferences account fails to recognise that individuals' preference for norm compliance are context-dependent, and so it ignores "a mapping from contexts to preferences that indicates in predictable ways how and why a given context or situation changes one's preference" (p. 581). Fehr and Rockenbach (2003) also overlook conditionality in their experimental study on altruism.

Bicchieri and co-authors implemented a number of experiments to investigate the joint effects and associations between descriptive and normative expectations. In this chapter I will first provide a detailed review of her philosophical conceptualisation of social norms. I will then raise a controversy implied by her conceptual analysis of social norms, and provide an alternative way to interpret her general proposal.

Expectations and conditionality are closely interrelated concepts in Bicchieri's definition of social norms. Bicchieri's theory of conditional norms *maps* specific contexts onto behavioural rules, and groups of such rules, taken together, give meaning to concepts such as fairness or trustworthiness.

This chapter will review Bicchieri's conceptual analysis of social norms. Given the purpose of this thesis is to investigate the operationalisability of Bicchieri's theory, the theoretical analysis in this chapter is established as

[4] This is the model, described in Chapter 1, which captures the idea that people dislike unequal outcomes, and they lose utility proportional to a distance measure of inequality.

operational concepts, instead of a focus of its philosophical clarity. The structure of this chapter follows the three kinds of condition which are suggested by Bicchieri for operational purpose. Given the key role of conditionality, this thesis calls Bicchieri's analysis a theory of conditional norms.

2.1 Conditionality and Expectations

2.1.1 Social Expectations

A social expectation is the type of expectation which can support the existence of social norms, according to Bicchieri (2006, 2017). It is the first key concept in Bicchieri's conceptualisation of social norms. They are *social* in the sense that they are based on what we believe about prevalence and distributions of others' behaviours and beliefs.

Bicchieri's model distinguishes between empirical expectations and normative social expectations.

Empirical expectations are what people think "others *will* do". Empirical expectations can support a social norm if *enough* other people in a certain population follow the norm in question. Empirical expectations are the first important basis for norm compliance. In a social interaction, in order for a person to play her equilibrium strategy she has to predict what the others' strategies will be. The definition of "enough", according to Bicchieri, includes two aspects. First, different individuals may have different thresholds below which they consider the number of followers as too small to count. Second, same individuals may have different thresholds for different norms (Bicchieri 2010).

A normative expectation is a belief about what behaviour, in a situation relevant to a norm, people in the population approve or disapprove of. This is usually expressed as "one ought to do X in situation Y". Normative expectations are the second important basis for norm compliance (Sugden 1998, 2004; Bicchieri 2006). The ground of normative expectations is that we expect others' beliefs about

what is praiseworthy or blameworthy that we have learned in the past to be continued in the future in the same situations.

An expectation is a type of belief. It is a belief about "what is going to happen or what should happen; both presuppose a continuity between past and future" (Bicchieri 2017, p. 11). Hence, in Bicchieri's analysis, following the belief/desire model, social expectations are modelled by social beliefs. In addition to distinguishing between empirical and normative social expectations, Bicchieri's model also distinguishes between *first-order* and *second-order* expectations/beliefs. First-order beliefs are regular beliefs, and second-order beliefs are beliefs about beliefs.

First-order social beliefs are *empirical* in the sense that one might believe/expect people to act in a certain way in some situations. I call these type of beliefs first-order descriptive beliefs (hereafter 1D). *Second-order* social beliefs are *empirical* when one believes that others in her society hold some first-order empirical beliefs about what others and herself are going to do in some situations. These types of beliefs are called second-order descriptive beliefs (hereafter 2D).

First-order social beliefs are *normative*, when one believes that others *should* behave in a certain way in some situations. *Second-order* social beliefs are *normative* when one believes that others in her reference network hold the first-order normative beliefs about how others and herself *should* behave in some situations. These two types of beliefs are called first-order normative beliefs (hereafter 1N) and second-order normative beliefs (hereafter 2N), respectively.

Therefore, the complete range of types of *social beliefs* that count as the social expectations which matter to Bicchieri's analysis of social norms are as follows (paraphrased from Table 2.1, Bicchieri 2017, p. 70):

First-order Descriptive belief (1D): what *one believes* about what *others* do.

Second-order Descriptive belief (2D): what *one believes* about what *others believe I/others* do.

First-order Normative belief (1N): what *one believes* about what others should do.

Second-order Normative belief (2N): what *one believes* about what *others believe I/others should* do.

Empirical expectations and normative expectations are the basis for a social norm when a set of conditions is satisfied, according to Bicchieri's analysis.

2.1.2 Co-Existence of Four Kinds of Social Belief

The initial formal statement of the conditionality of social norms given by Bicchieri (2006, p. 11) is as follows:

> *"Conditions for a Social Norm to Exist*
> Let *R* be a *behavioral rule* for situations of type *S*, where *S* can be represented as a mixed-motive game. We say that *R* is a social norm in a population *P* if there exists a sufficiently large subset $P_{Cf} \in P$ such that, for each individual $i \in P_{Cf}$:
> (1) *Contingency*: i knows that a rule *R* exists and applies to situations of type *S*;
> *Conditional preference:* i prefers to conform to *R* in situations of type *S* on the condition that:
> (2) (2a) *Empirical Expectations*: i believes that a sufficiently large subset of *P* conforms to *R* in situations of type *S*;
> and either
> (3) (2b) *Normative Expectations*: i believes that a sufficiently large subset of *P* expects i to conform to *R* in situations of type *S*;
> or
> (4) (2b') *Normative Expectations with Sanctions*: i believes that a sufficiently large subset of *P* expects i to conform to *R* in situations of type *S*, prefers i to conform, and may sanction behavior.
> A social norm *R* is *followed* by population *P* if there exists a sufficiently large subset $P_f \in P_{Cf}$ such that, for each individual $i \in P_f$, conditions (2a) and either (2b) or (2b') are met for i and, as a result, i prefers to conform to *R* in situations of type *S*".

This statement shows the first type of condition for an individual's preference for conforming to a social norm, according to Bicchieri. It is the co-existence of the set of social beliefs, i.e., 1D, 2D, 1N, and 2N beliefs. Following Bicchieri's language, I call them a set of four conditions. Table 2.1 below shows this set of conditions

which covers the full range of the four types of social beliefs, regarded as social expectations in Bicchieri's analysis. This table is adapted from Table 2.1 in Bicchieri (2017, p. 70).

Table 2.1 Full set of social beliefs required in Bicchieri's theory of social norm

Social belief	First-order	Second-order
Descriptive	1D (Condition 1) what *one believes* about what *others* do.	2D (Condition 2) what *one believes* about what *others believe I/others* do.
Normative	1N (Condition 3) what *one believes* about what others should do.	2N (Condition 4) what *one believes* about what *others believe I/others should* do.

Next, I will explain these *four* conditions one by one, relating to some of Bicchieri's original texts. Along with the specification of each condition, I will explain why each of them is needed for norm identification according to Bicchieri. I argue these four conditions are what constitute the *conditional norm* as analysed by Bicchieri. I further argue that an identification of the existence of *conditional norm* requires *all* four types of beliefs. In this section, I take the concept of belief and expectation as interchangeable concepts.

Condition 1 & Condition 2 – Empirical Expectations

Condition 1 and condition 2 both refer to empirical expectations. However, they refer to different types of empirical expectations. Condition 1 refers to the first-order empirical expectations which are picked up by 1D beliefs; and condition 2 refers to the second-order empirical expectations which are picked up by 2D beliefs.

A 1D belief is what one believes about what others actually *do*. Condition 1 as 1D belief can be first seen from Bicchieri's explanation of "contingency" that "the *contingency condition*, says that actors are aware that a certain behavioural rule

exists and applies to situations of type *S*. This collective awareness is constitutive of its very existence as a norm" (2006, p. 12). Condition 1 as 1D belief can also be seen from Bicchieri (2006, p. 13) which argues that "the empirical expectations condition, says that expectations of *conformity* matter. … Such expectations are taken to be empirical expectations in the sense that one expects people to follow R in situations of type S because one has observed them to do just that over a long period of time".

The second condition refers to the second-order empirical expectation, that is the 2D belief. Condition 1 and condition 2 are differentiated by the concept of *conformity*. Bicchieri (2006, p. 112) argues that "to be effective, norms have to be activated by salient cues…. A norm may exist, but it may not be followed simply because the relevant expectations are not there, or because one might be unaware of being in a situation to which the norm applies. Here the conditional preference for *conformity* refers to two senses of expectations that (a) agents expect others to follow it and (b) they believe that, in turn, they are expected by others to abide by the norm". In this statement, (a) refers to the *contingency* condition which are referred to as the 1D beliefs about what one believes others will do, whereas (b) refers to a 2D beliefs about the 1D beliefs. Therefore, according to Bicchieri, for an identification of the conditional norm, we need the 1D beliefs to test the *contingency* condition; and the 2D beliefs to test for *conformity*. This suggests a joint elicitation of 1D and 2D beliefs.

It is possible that Bicchieri might have never clearly distinguished the 1D and 2D beliefs in her theoretical work (e.g., Bicchieri 2006). The need to elicit 2D beliefs for norm identification is made explicit later in her empirical work (see Bicchieri 2017, p. 70, footnote 11). In fact, the expression "empirical expectations" in Bicchieri's empirical work usually refers to just 1D beliefs (Bicchieri and Chavez 2010, 2013; Xiao and Bicchieri 2010; Bicchieri and Dimant 2019). However, in this thesis, I maintain a clear distinction between condition 1 and condition 2. I further suggest that applications of her theory should draw clear distinctions between these

two conditions, particularly for a consideration of an operational purpose.

Condition 3 & Condition 4 – Normative Expectations

Condition 3 and condition 4 refer to normative expectations.

Condition 3 refers to the first-order normative expectations which are picked up by 1N beliefs. In Bicchieri's analysis, she calls this type of belief *personal normative beliefs*. Bicchieri further distinguishes *personal normative beliefs* into two kinds. The first kind is beliefs about "what *I* should do", they are the first-order *personal* normative beliefs in the strict sense. The second type is beliefs about "what *others* should do", they are the first-order *social* normative beliefs. Given it is social dimension which my thesis is interested in, this thesis leaves aside the conceptual difference Bicchieri draws about the 1N belief. Hence, in this thesis my application of 1N belief only refers to the second type, i.e., "what one believes *others* should do".

Condition 4 refers to the second-order normative expectations, which are picked up by 2N beliefs.

Bicchieri's analysis suggests a joint elicitation of both 1N and 2N beliefs.

On the one hand, Bicchieri (2017, p. 70) argues that 1N beliefs need to be elicited in order to assess the mutual consistency of normative beliefs, between 1N and 2N. The mutual consistency between these two types of normative beliefs arguably allows one to determine the degree to which a behaviour is endorsed. For example, the experimental study in Bicchieri (2022) posed a 1N question to their subjects about child marriage: "Some girls get married before they are 18 years old, is this good?". Meanwhile, the same study also claimed that they asked a 2N belief question that "Do you think people in your community believe that it is a father's duty to marry off daughters as soon as possible?". Bicchieri (2022) argues that if the 1N beliefs and the 2N beliefs do not match, it would become impossible to evaluate the accuracy of the normative expectations (2N beliefs). This suggests a joint elicitation of 1N and 2N beliefs.

On the other hand, Bicchieri argues that the importance of eliciting 2N beliefs jointly with the 1N beliefs, is that 2N beliefs are needed to test the 1N beliefs. 2N beliefs, according to Bicchieri, provide crucial evidence that behaviour might be governed by a norm. Bicchieri (2006, p. 15) argues that people may have different reasons for conditionally preferring to follow a norm. If an agent unconditionally behaves in accordance with a shared belief, it is most likely that the agent applies a personal normative belief, e.g., a prudential norm.[5] In this case, practical reasoning might weigh heavier than a social constraint. Whereas, if an agent appeals to a shared belief (attitude) due to social expectations, per Bicchieri's analysis, then it is likely that practical reasoning might play less weight in motivating the agent to comply with a certain norm.

Similar to the ambiguity between condition 1 and condition 2 in Bicchieri's analysis, there is also conceptual ambiguity between condition 3 and condition 4. In Bicchieri's empirical work, the term "normative expectations" only refers to the 2N beliefs, and it seems that condition 3 is not directly included in Bicchieri's definition of social norms. However, Bicchieri addresses the importance of 1N beliefs for experimental settings for norm identification. Bicchieri also made it clear that "to identify social norms, we usually only need to gauge personal normative beliefs and empirical/normative expectations…" (2017, p. 71). Again, in this thesis, I maintain a clear distinction between condition 3 and condition 4, and I suggest the distinction should be drawn particularly in operational applications of her theory.

Nevertheless, the ambiguity I pointed out here may imply that Bicchieri's empirical research may suggest the 1D beliefs and 2N beliefs carry significant weight for norm identification, whereas the other two types of beliefs (i.e., 2D and 1N) carry less weight. For example, in Bicchieri and Chavez (2010) 1D and 2N

[5] A prudential norm is what agents comply with due to prudential reasoning (prudential belief); it is regarded as unconditional or less conditional. In Bicchieri's analysis, she disregards prudential and moral norms in her theory of conditional norms, although she claims that they are *social* in the weak sense (2017, p. 71). The "weak sense" is what I mean here by "less conditional".

beliefs are elicited, in Xiao and Bicchieri (2010) 1D and 1N beliefs are elicited, and in Bicchieri and Dimant (2019) 1D and 2N beliefs are elicited. However, I maintain this to be an empirical question, not a conceptual one. I do not deny the possibility that some kind of beliefs may carry less weight depending on the empirical circumstances. This conjectured fact is captured in Kuran's (1995) model of social influences on utility. His model accommodates this fact by attaching weighting parameters to different types of beliefs. Bicchieri's model has no comparable resource. Hence there is no basis in applying Bicchieri's conceptual framework for social norms to exclude *a priori* any one of the four belief frames. I argue that the clarity about the need for measurements of all four types of beliefs at the conceptual level must be made explicit.

To conclude, the set of four social beliefs constitutes the first set of conditions in Bicchieri's analysis of social norms. In the concordance with Table 2.1, these four types of expectations are picked up by four types of *social* beliefs which are referred to in this thesis as 1D, 2D, 1N, and 2N. In the experimental study for this thesis, we test the existence of *all* four types of beliefs, i.e., 1D, 2D, 1N, and 2N beliefs.

2.2 Belief Consistency

Corresponding to the set of four conditions, another key requirement of Bicchieri's theory of conditional norms is *belief consistency*. However, there is some ambiguity about what belief consistency means. I conjecture that in Bicchieri's analysis there should be two types of relevant belief consistency. The first type is *within-individual* belief consistency. Within-individual belief consistency concerns the degree of consistency between different (but relevant) beliefs of an individual. The second type refers to belief consistency *across individuals*. Across-individual belief consistency concerns the degree of consistency of one belief across different individuals in a population.

This section will review these types of belief consistency, and the implications

for statistical tests of observed choice behaviour to identify norms.

2.2.1 Two Types of Consistency

The first type of belief consistency is *within-individual* consistency. For example, Bicchieri (2006, 2017) argues that an agent's norm compliance is crucially dependent upon the *co-existence* of empirical expectations and normative expectations. To be specific, this refers to a *within-individual* belief consistency, the consistency between what an individual thinks others will do (1D belief) and what she thinks others think should be done (2N belief). Bicchieri (2017, p.70) points out that when there is *inconsistency* between normative and empirical expectations (2N versus 1N), we should not identify a norm. Sometimes, this sense of consistency is referred as *congruence*. For example, Bicchieri and Dimant (2019) argue that "if empirical and normative expectations are *incongruent*, a norm will not be obeyed" (p. 5).

Consistency across individuals is what Bicchieri calls mutual consistency or *consensus*. For example, Bicchieri (2017) argues that "when a social norm exists and applies to a specific situation, the second-order *normative expectations* (between different individuals) will be mutually consistent, if these expectations are mutually consistent, then there is widespread *consensus* that a specific behavior should be performed (or avoided)" (p. 70). In other words, "consensus means an agreement in *individuals' second-order normative expectations*, that is the presence of a shared norm" (Bicchieri and Chavez 2010, p. 163). This refers to the *across-individual* consistency of the 2N beliefs.

The two types of belief consistency are not exclusive. In fact, Bicchieri suggests an identification of both kinds of belief consistency. For example, Bicchieri (2017) argues that to identify a norm, it needs to be established primarily that there is a consensus about what actions are appropriate/inappropriate in what situations, that is the measurement of *across-individual* belief consistency of the 2N beliefs.

However, Bicchieri also argues that “when second-order beliefs (normative) are mutually consistent *but systematically inaccurate*, we know that people uphold a norm they dislike. If their doubts are not shared, the norm will persist” (2017, p. 74). This is to say, norm identification requires a measurement of within-individual consistency between the 1N and 2N beliefs, in addition to the existence of *across-individual* consistency of the 2N beliefs.

However, what does it mean to say that beliefs are consistent? How similar must they be to be counted as consistent? From a statistical point of view, how much overlap between two belief distributions satisfies a threshold for consistency? In other words, what is required with respect to estimation of the mean or the mode, and a standard deviation of two belief distributions when we claim them to be consistent?

The notion of similarity has been analysed from a feature-theoretical approach by Tversky (1977), which challenges the previous domination of geometric models. Geometric models measure the similarity relation as the metric distance between points in coordination space. However, a feature-theoretical approach defines similarity between objects as the value of a feature-matching function. Based on quantitative assumptions about similarity ordering, the similarity of two objects is characterised as a linear combination (or contrast) of the measures of their common and distinctive features (Tversky and Gati 1978). Similarity increases with the measure of the commonality and decreases with the measure of the distinction. Tversky and Gati (1978) employed the contrast model to analyse three empirical problems. The results suggest that the parameters of the contrast model are sensitive to manipulations.

In Bicchieri’s theory of social norms, belief consistency/similarity is applied as an independent variable to explain norm compliance. However, the concept of belief consistency is left as a practical question, not a conceptual one.

2.2.2 Statistical Insights

The concept of belief consistency provides major leverage for statistical insights in empirical studies of social norms. However, given the difference between the two types of belief consistency specified previously, we must clarify the following three issues for the purpose of statistical modelling of empirical belief consistency.

First, what types of beliefs are relevant to the statistical comparison of belief consistency? The conditionality of norm-relevant preferences suggests that belief consistency must be measured between *all* four types of beliefs.

Bicchieri (2017, p. 70, Footnote 11) writes that:

> "To identify social norms, we usually only need to gauge personal normative beliefs and empirical/normative expectations, but on occasion we may also want to evaluate second-order empirical expectations, as when we ask people not just what they or others would do, but also what they think others believe people in their situation would (as opposed to should) do".

This may seem to suggest that occasionally there is no need to elicit 2D beliefs at all. However, I argue that this is mistaken given Bicchieri's conceptual analysis of social norms. A key issue is that, as specified in Bicchieri (2006, p. 22), a social norm requires that the action suggested by the social norm be an equilibrium strategy. This is where the 2D and 2N conditions come in, to ensure that we have a unique Nash equilibrium conditional on actual beliefs. An equilibrium here means that when agents choose actions suggested by an effective social norm, their choices reach a "situation of stable mutual adjustment: Everyone anticipates everyone else's behaviour, and all these anticipations turn out to be correct. In other words, an equilibrium is a set of self-fulfilling prophecies that individuals formulate about each other's actions" (Bicchieri 2006, p. 22). Such an equilibrium provides a reference point for individuals' decision making, and allows for social coordination.

The second issue is the distinction between *overall* consistency and *partial* consistency. Let us define *overall* belief consistency as consistency between all four types of beliefs, i.e., first-order and second-order descriptive beliefs and first-order

and second-order normative beliefs.[6] *Partial* consistency refers to an "untransferable" consistency, for example both first-order normative beliefs and second-order descriptive beliefs are consistent with the first-order descriptive belief, however, the first-order normative beliefs are inconsistent with the second-order descriptive beliefs. In an experimental setting, tests of overall consistency can be extremely complex, tests of partial consistency can induce an inference about overall consistency. In our experiment, in order to investigate Bicchieri's conceptual analysis in the lab, we test partial consistencies.

The distinction between overall consistency and partial consistency opens the possibility of variations of degrees of consistency being identified by demographic variables that partition groups into different reference networks. The deviation is also relevant to *within-individual consistency*. Partial belief consistency within individuals is what causes norm uncertainty, according to Bicchieri. Statistical modelling of belief inconsistency in this sense should be able to explain what degree of consistency is required for norm compliance, and what degrees of inconsistency can explain social phenomena such as the existence of norm disruptors, norm manipulation, and pluralistic ignorance. For example, Dana, Weber & Kuang (2007) conducted an experimental study in which they gave subjects moral "wiggle room" in interpreting norm compliance and found that subjects given such scope for rationalisation behaved more self-interestedly. Tracking inconsistency between empirical and normative beliefs in similar experimental settings may yield inferences which explain why fairness is a social norm rather than a set of social preferences.

With respect to *across-individual consistency*, we have to be clear about what is required in statistical modelling for capturing precisely the hierarchical structure between beliefs at the individual level and beliefs at the group level. Bicchieri seems to suggest (in various places) that when individuals' normative expectations align

[6] Bicchieri (2006) isn't committed to such analysis in general.

with **what *the majority*** in their reference network approve or disapprove of, this counts as consistency between the belief at the level of individual and belief at the group level. However, what counts as a "majority" can differ between different people and differ in application to different social norms. The statistical modelling of belief consistency at the group level must be able to capture what counts as a majority with respect to some specific norm, or regarding norms in general. In Bicchieri's analysis, she has been vague about how much of a population has to expect conformity to a norm to count as "enough" for the existence of a norm. Indeed, as Bicchieri points out (2017, p. 108) what counts as a "majority" varies from case to case. In this thesis, we set this issue aside.

The third issue we must consider is the scale of the comparison of belief consistency. This is related to the concept of a reference network. Reference networks define sub-group norms, because reference networks determine normative expectations and what actions and attitudes people care about.[7] What constitutes a reference network is the interactions people have among each other in a certain community.

In the experimental study for this thesis, we test the primary condition with respect to the concept of belief consistency, that is the *overall* belief consistency across *all four* types of beliefs, measured at the *group level*. In practice, a test of an overall consistency between four types of belief can allow various levels of partial consistency. For example, in our experiment in order to test the consistency between 1D, 2D, 1N, and 2N beliefs, we start with testing the partial consistency between 1D and 1N, and between 2D and 2N. This allows us to infer the degree of overall consistency.

[7] Section 2.3 provides a more detailed review of the concept of reference network.

2.3 Reference Network

2.3.1 Defining a Reference Network

Reference network is the third condition suggested by Bicchieri's theory of conditional norms. Bicchieri (2017, p. 11) characterises a reference network N_i as the set of people whose normative expectations matter to the behaviour and normative beliefs of an individual agent i. A reference network is a cluster within a larger social influence network (Goyal, 2007).[8] "Clustering" here simply means that density of mindshaping relationships[9] is high within the cluster relative to other parts of the larger network. Reference networks may involve heterogeneity in any of these relevant dimensions. There are three features about the heterogeneity of reference networks. The first concerns the scale: a reference network can be local or general, and be big or small. The second dimension is variation in conditions under which a reference network is at play in an agent's decision making. The third dimension is the dynamics: agents' reference networks can change over time. The concept of a reference network reflects the characteristic that a norm is a characteristic of a community. Communities can be specified relevantly broadly or narrowly.

According to Bicchieri's analysis, reference networks matter for shaping agents' conditional preferences, particularly the 2N beliefs. Bicchieri (2017, p. 53) warns that when the reference network is not taken into account in a normative intervention, the intervention is likely to fail. One example is a case study of a failed normative intervention to promote breastfeeding in West and Central Africa. In the case study, researchers (Ayoya et al. 2012) attempted to change behaviours of new mothers, to encourage them to breastfeed their babies. The intervention consisted of information provision. Behaviours of mothers did not significantly change as a result. The failure arguably stemmed from the fact that mothers-in-law and older women in general, who play a key role in young mothers' reference networks, weren't included

[8] See Chapter 2 of Goyal (2007) for the relevant concept specification.
[9] See Zawidzki (2013).

in the intervention. Therefore, unchanged social expectations, especially the normative expectations from new mothers' reference networks, still were the main determinants that overwhelmed any effects of the information intervention on beliefs.

The heterogeneity of reference networks also reflects the fact that individuals live in interacting social webs. It is normal that an agent's decision making about different conditions may engage influences from different reference networks: depending on different circumstances, different people will matter differently to decisions. Different reference networks may have different normative expectations, and it is only the normative expectations from the people who matter to a specific decision that will motivate norm-compliance or norm-violation. Bicchieri (2017, p. 14) gives an example of a Pakistani immigrant to Italy who killed his "dishonored" daughter even though he had lived in Milan for 20 years. The social norms prevailing in Milan and the normative expectations of his friends and colleagues in Milan weren't as effective in influencing him as the norm of family honour prevalent in his Pakistani village. The family and people in the village played the role of the reference network governing his response to his daughter's perceived transgression.

Reference networks are dynamic. Muldoon (2022) argues that when reference networks split, some sub-groups stop following previous norms. This suggests that one potential approach to norm intervention is to alter reference networks, which in turn can change agents' social expectations, especially normative expectations. For example, Bicchieri (2017, p. 158) suggests that if deliberation were to be applied to changing a norm, the deliberation promoted should not be restricted to the population at risk (e.g., young women in the case above), but it should include all the members of the reference network of the targeted subgroup. The reference network may include men, relatives, extended families, religious leaders, village leaders, and others.

2.3.2 Methodological and Statistical Insights

Norm modelling must be sensitive to reference networks in the sense that when we investigate a social norm, we must specify which community the investigation refers to. Bicchieri (2017, p. 53) argues that "mapping the reference network is essential part of understanding social norms and how to change them, because the norm has to change within the reference network".

Therefore, an empirical study investigating social norms must clearly define which reference network is referred to when eliciting beliefs. Reference networks can sometimes be identified through demographic and/or normative surveys. Information about reference networks given to subjects might induce them to respond in that reference network frame when they report their beliefs. In the statistical modelling procedure, such information will be reflected in model covariates. Empirical studies might measure the degree of consistency of different reference networks through varying demographics in samples and across experimental treatments.

With respect to the scale of a reference network, what matters to empirical studies is agents' *confidence levels* in their beliefs. Suppose a study randomly sampled subjects across a very large community, across a nation for example. We might expect that any social norm we found would be reflected in low confidence across subjects' beliefs. This is because subjects might anticipate significant heterogeneity across a whole nation. This would be expected to lead to diffuse prior beliefs. By contrast, if a study used a highly cohesive community sample with a tight reference network, and if subjects were aware of the reference network, this would be expected to lead to higher confidence in beliefs.

Another feature to consider is the dynamics of the reference network. In studies where experiments run across a period of time, experimenters should allow for the possibility of shifting reference networks, reflected in subjects' beliefs over the study's time course.

One way to consider reference networks in the modelling of belief consistency is through collecting and controlling for information on features of individuals such as gender, age, and education. Normative attitudes count as another type of information.

In the experiment conducted for this thesis, we control for the reference network alignment by administering a demographic questionnaire and a normative values survey. This allows us to collect some exogenous information which we conjecture might allow for identification of relevant reference networks.

2.4 A Philosophical Reflection

By contrast to conditional norms, Bicchieri's analysis also introduces a concept of unconditional norms. Unconditional norms refer to the kind of norms to which people's motivation for compliance is unconditional, Bicchieri calls them "personal" norms. For example, prudential and moral norms are counted as unconditional in Bicchieri's analysis. An example of behavioural compliance to a prudential norm, given by Bicchieri, might be that an agent follows a non-smoking norm and refuses to smoke because she believes that smoking hurts her and others' health. By contrast, if an agent follows a non-smoking norm because she believes that women are "debauched" if they smoke, then this is behavioural compliance to a moral norm.

The distinction between conditional and unconditional expectations mainly sets the scope within which Bicchieri's theory applies. However, I suspect this confronts her theory with a serious "scope issue".

The main problem is that the understanding of moral beliefs being unconditional is not sustainable. First, it is philosophically implausible that a moral norm could be completely personal and exist apart from the attitudes of any others in the society. It is also highly doubtful that many other philosophers would agree with Bicchieri's account to regard moral beliefs as unconditional. For naturalistic philosophers such as Richard Rorty, Donald Davidson, Daniel Dennett, Andy Clark,

Don Ross, moral beliefs are also conditional: moral beliefs arise from the same process of socialisation (i.e., learning, imitation, pedagogy) as other norms. Furthermore, seeing moral norms as personal disengages with the empirical reality about how the social world can shape individuals' moral beliefs (Wolf and Koons, 2016; Burge, 1986; Sneddon, 2011; Binmore 1994, 1998). It is also highly doubtful that other theories of social norms can provide empirical support to how moral norms can be excluded from social norms. For example, many social issues modelled by Kuran (1995) as preference falsification depend on features peculiar to *moralised* norms.

However, given that the main objective of this thesis is to investigate the operationalizability of Bicchieri's concept of a conditional norm, I will postpone further attention to the consequences of her exclusion of moral norms for the policy significance of her theory until Chapter 7.

2.5 Summary

To conclude, I reviewed Bicchieri's theory of conditional norms as a basis for methodology for the empirical study of social norms. I reviewed the three kinds of conditions involved in Bicchieri's analysis. The three conditions for the existence of a social norm are: co-existence of the four kinds of social beliefs; consistency between the set of beliefs; and an alignment between the reference networks. Corresponding to each condition, I also suggested the hypothesis to be tested in our experimental study, with respect to each condition. I also provided a reflection on a "scope issue" for Bicchieri's theory given her understanding of moral norms being excluded from the theory of conditional norm.

One question might be whether a test of Bicchieri's theory of conditional norms needs complete overall consistency across all four types of beliefs within one reference network. This thesis concludes as follows: 1D and 2N belief alignment is obviously needed. Now suppose that 2D beliefs were unaligned. People would then

expect others to conform to and endorse a norm but not expect others to expect conformity. So we would have to suppose that they all think there is a norm but that most people don't know that it regulates behaviour. That is *conceivable,* but it would be very strange, in the sense that everyone would think that everyone else was a zombie about this norm. Now, suppose that 1N beliefs are unaligned. In this case everyone expects everyone to expect endorsement of the norm but not to actually endorse it. This is also *conceivable,* as in a statement such as "they all expect each other to (e.g.) promote bigotry but they also expect everyone to deny that they themselves are bigots". This example shows that it seems that the case could only apply to a norm people were embarrassed about. One could say, yes, this is possible, but only in a very cynical society with very low trust. There would be some obvious strain in claiming that this society operated a norm against bigotry.

Hence, I conclude that the standard empirical case applying Bicchieri's theory of conditional norms should test complete consistency across *all* four types of beliefs.

Therefore, the primary hypothesis for the experimental study in this thesis is to test whether there exists an *overall* belief consistency across *all four* types of social belief (1D, 2D, 1N and 2N) within one reference network. Statistically, this will be tested at the group level.

Chapter 3 - Philosophical Frames for Bicchieri's Theory of Conditional Norms

As reviewed through sections 2.1 to 2.3, intentional concepts such as beliefs, expectations, and preferences are the building blocks for Bicchieri's conceptualisation of social norms. However, it is not clear in her conceptual analysis what these concepts refer to. This chapter proposes the Dennettian account of the Intentional Stance (IS) as the philosophical framework for the intentional concepts applied in CB's conceptual analysis of social norms. After a brief sketch of Revealed Preference Theory (RPT), I then argue that both the IS and RPT can be synthesised to construct an understanding of how the intentional concepts applied in Bicchieri's analysis of social norms can be measured and modelled for empirical purposes.

3.1 The Intentional Stance

Beliefs are representations about how the world is. In the tradition of philosophy of mind, mind-to-world fit mapping typically interprets beliefs as "the mind fitting the world", which means beliefs reflect how the world is, and desires as "the world fitting the mind". Beliefs are intentional, that is, about something. In everyday life, we often appeal to others' propositional attitudes to explain their actions and/or predict their further behaviour. This is what Dennett calls "taking the intentional stance". The designation *propositional attitudes* characterises the family of relationships between agents and intentional content. According to Dennett (1991), propositional attitudes such as beliefs and desires are abstract posits that help us to track real patterns in observable behaviour, including linguistic behaviour. The IS is one of the three predictive and explanatory strategies[10] one can adopt for explanation of complex systems (Dennett 1987). To adopt the IS is to treat the system as if it had

[10] The other two stances which act as predictive and explanatory strategies, are the physical stance and the design stance (Dennett 1987, 1988, 1991).

motivating states characterised by propositional attitudes like belief, desire and preference which are about some object, situation, or event (Dennett 1987, 1988, 1991).

According to Dennett, the intentional stance identifies 'true believers'. A true believer is a physical system whose behaviour can only adequately be explained using the IS. To treat someone as an intentional system, and therefore, as a true believer, is to treat them as rational, in a sense that Zawidzki (2007, p. 48) formulates as follows:

> "(1) Intentional systems have the beliefs they ought to have, that is, true and relevant beliefs, given their perceptual capacities and informational needs; (2) Intentional systems have the desires they ought to have, given their biological needs; (3) Intentional systems behave in ways that count as rational given these beliefs and desires".

The normativity ("*ought to*") implies that to attribute propositional attitudes according to the IS is not to conjecture private mental states that are causally responsible for behaviour. Rather, it is to locate behaviour in a normative frame to allow the behaviour to count as reasonable given the agent's goals and available information (Zawidzki 2012, 2013). Such rationality is not a fiction or a hidden complex within an agent's head, but it is socially constructed. The concept of rationality in the Dennettian sense is characterised as "ecological rationality" by the economist Vernon Smith (2007). Ecological rationality refers to "emergent order in the form of the practices, norms, and evolving institutional rules governing action by individuals that are part of our cultural and biological heritage and are created by human interactions, but not by conscious human design" (p. 2).

The concept of rationality in Dennett's and Smith's sense is supported by the externalist account of the mind. Cognitive externalism (Clark 1997, 2008) challenges the internalist account in respect of the causal relation between mental states and behaviour. It claims that propositional attitudes are not psychological states which are private in individuals but are social constructs. Propositional attitude ascriptions do not project causal influences from mental states onto

behaviour, but pick out relations among an agent, features of their environment, and patterns of her social expectations. Ross (2014) argues that "the mind itself, as understood by externalists, is an interface pattern that describes the systematic relationships between brains and socially related people" (p. 242).

Belief in this sense is abstract, virtual, and based on social facts (Gilbert 1989). Ross (2014) promotes Dennett's notion of realism in his articulation of intentional-stance functionalism, according to which "belief is not an occurrent state in the brain, it is a relation between the brain and a social environment structured by norms and mutual expectations. Thus, the belief is a kind of virtual state. Being virtual is not a way of being fictitious; it is a way of being real. If one needs to track a pattern in order to explain and predict actual events, that pattern is real" (p. 248).

The two main rivals of the IS are computationalism and eliminativism. Both computationalism and eliminativism claim that the IS is a form of instrumentalism. Instrumentalism is the view that the propositional attitude ascription is metaphorical, and that the IS is nothing but a useful epistemological tool, an instrument for generating predictions. For example, Fodor (1996) argues that Dennett's account does not tell us anything about what propositional attitudes are and how to differentiate one mental state from another.

To respond to the critiques, Dennett (1991) developed the concept of "real pattern" and counters that ignoring attributed intentionality would lead to missing real patterns.

Dennett argues that the intentional stance is a way of seeing agency in a broad variety of systems, i.e., thermostats, computers, simple organisms, and even natural selection. For some systems, the IS is an optional strategy adopted for convenience, but for some others it is required by a full account of reality. The distinction given by Dennett (1987) between a house thermostat and a chess-playing AI is his leading example. A thermostat lowers the temperature in the house when it exceeds a threshold. We might apply the IS to the behaviour by ascribing the propositional

attitude that the thermostat *prefers* the room temperature to stay at the threshold. In the case of a chess-playing AI, we might apply the intentional stance by stating that the computer *wants* to get its queen out early in the game. What is important is that we can but do not *have to* assume the IS towards the thermostat (because the physical stance would be enough for explaining and predicting its behaviour). However, we *must* adopt the IS towards the chess-playing AI because a lower-level stance such as physical stance or design stance won't be enough for the purpose of predicting what the computer will do next. Such prediction requires attention to what constitutes competent chess strategy.

Although Dennett (1991) illustrates another example of a real pattern by appealing to the Game of Life, Ross (2000, 2005) maintains that this example fails to demonstrate the reality of the patterns observed. Ross (2000) argues that Dennett's example of the Game of Life does not adequately capture real patternhood because in Game of Life the global pattern/behaviour can be completely derived from finite rules governing individual cells. Other kinds of machines, e.g., connectionist machines that utilise deep learning neural networks, exhibit global patterns of behaviour (e.g., chess-playing behaviour) that cannot be completely inferred from the rules governing individual neurons or subsystems.

Ladyman and Ross (2007) complete Dennett's theory of real patterns and provide a technical analysis of real patternhood in terms of computational efficiency and non-redundancy of projectability. This allows the argument against instrumentalism about propositional attitudes according to the IS is to be made complete.

The IS therefore is not simply a strategy for taking an epistemological shortcut to apprehending minds. It is the stronger suggestion that all there is to having a belief that *p* is being a system that is predictable under the assumption that it believes *p*. This suggestion carries both ontological and methodological weight. Ontologically, to say that *x* believes *p* is to assert that *x*'s behaviour (verbal and

otherwise) demonstrates a particular kind of regularity; namely, just that kind of regularity which justifies the 'projection' (the attribution by the observer) of *x*'s attitudes to the proposition that *p*. Methodologically, the propositional attitudes such as beliefs and desires are virtual properties of agents in environmental (particularly social) contexts which can be attributed from the IS.

The emphasis on expectations and conditionality in Bicchieri's theory of social norms is compatible with the intentional stance in a sense we can render as follows: Bicchieri's account of social norms suggests that the relevant networks of expectations are social facts the agent encounters and that the best methodology for measuring behavioural responses to these social facts is to construct utility functions from choice behaviour that are conditional on recognition of any norm, but do not vary with outcomes distinguished by reference to their normative assessments. The IS is the philosophy of mind that supports this conception, according to which the propositional attitudes, such as beliefs and desires, are virtual properties of agents in environmental (particularly social) contexts.

3.2 Revealed Preference Theory

In economics, Samuelson's Revealed Preference Theory (RPT) showed the necessary mathematical conditions which enable ascription of a consistent set of preference to an agent based on her behaviour without taking a position about underlying psychology (Samuelson 1948). As Hands (2013) argues, Contemporary Revealed Preference Theory (CRPT), as an extension of the original RPT, not only allows for the construction of utility functions empirically from finite choice data, but also embeds philosophical and methodological commitments. CRPT, like other versions of RPT, defines preference solely in terms of choice behaviour and denies that preferences are causes of choices. It advocates interpreting RPT as a general methodological template for all of microeconomics. My arguments about RPT mainly refers to CRPT but includes RPT in general terms.

Binmore (2009) points out that the RPT "makes a virtue of assuming nothing whatever about the psychological causes of our choice behaviour" (p. 8). This means an RPT analysis is based only on information about what subjects choose. Given observed behaviour tells us people choose in some situations, these data can be used to deduce what they would choose in other situations. However, in making such inferences there are two assumptions about agents' choice behaviour: stability and consistency (Binmore 2009, p. 9-12).

Stability and consistency allow the use of RPT to model agents as seeking to maximise the value of something. Whatever this abstract something may be in a particular context, economists call it "utility". The utility function constructed based on assumptions of stability and consistency therefore can be synthesised as the real pattern tracked by our intentional states from observable behaviour according to the IS.

Belief attribution according to the IS requires a similarly thin assumption of rationality, according to which there is a systematic relationship between agents' choices and their goals – a real pattern. This sense of rationality is called "ecological rationality" in economics, and "normativity" by Zawidzki (2007). What merits emphases about such synthesis between the RPT and the IS are two points. First, neither theory aims to explain rationality, i.e., neither the RPT nor the IS identifies some preferences or beliefs are more rational than others. Second, neither theory aims to identify causes of behaviour. The pattern does not rule out errors. It requires only stochastic dominance of choices that produce "better" outcomes in expectation.

In RPT, the only criteria that is relevant when picking one of the infinity of utility functions that represent a given preference relation is that of mathematical convenience. Von Neumann and Morgenstern's (VNM) approach to utility gives a primary role to risk, VNM's utility measures how much an agent wants something by the size of the risk she is willing to take to get it. Expected Utility Theory (EUT) requires that the preferences over lotteries that an agent reveals in risky situations be

consistent.

EUT was used by Savage (1972) to develop the standard model of subjective probability. According to EUT, an agent acts as though she believes that each possible state of the world in a decision problem has a subjective probability. Savage then models the agent's utility by assuming that there exists a subjective probability distribution *and* a utility function such that observed choices can be characterised as maximising her Subjective Expected Utility (SEU). The existence proof for identification of SEU requires satisfaction of the following axioms:

P.1 (Ordering) The relation ≤ is a simple ordering among acts.
If *F* is a finite set of acts, there exist *f* and *h* in *F* such that for all g in *F*.

$$f \leq g \leq h$$

This axiom defines that the preference relation as a complete, transitive and reflexive binary relation on *F.*

For the next axiom, where ~*B* is the complement of event *B,*

P.2 (Sure-Thing Principle) If acts f, g, and modified acts f', g' are such that:
1. in ~*B*, f agrees with g, and f' agrees with g',
2. in *B*, f agrees with f', and g agrees with g',
3. $f \leq g$;

then $f' \leq g'$.

This axiom is also called Sure-Thing Principle. Savage states that: "if the person would not prefer f to g, either knowing that the **event** *B* obtained, or knowing that ~*B* obtained, then he does not prefer f to g. Moreover (provided he does not regard *B* as virtually impossible) if he would definitely prefer g to f, knowing that *B* obtained, and, if he would not prefer g to f, knowing that *B* did not obtain, then he definitely prefers g to f" (p. 21- 22). The Sure-Thing Principle implies that preferences are separable across events, hence it maintains an interpretation of the consistency of dynamic actions.

P.3 (Monotonicity) If $f = g$, and $f' = g'$, and *B* is not null; then $f \leq f'$ given *B,* if and only if $g \leq g'$.

This axiom states that the preference between two acts are state-independent, and that the preference between two acts depend solely on the consequences in states in which the payoffs of the two acts being compared are distinct.

P.4 (Weak Comparative Probability) If f, f', g, g'; A, B; f_A, f_B, g_A, g_B are such that:

1. $f' < f$, $g' < g$;
2a. $f_A(s) = f$, $g_A(s) = g$ for $s \in A$,
$f_A(s) = f'$, $g_A(s) = g'$ for $s \in \sim A$;
2b. $f_B(s) = f$, $g_B(s) = g$ for $s \in B$,
$f_B(s) = f'$, $g_B(s) = g'$ for $s \in \sim B$;
3. $f_A \leq f_B$;

then $g_A \leq g_B$.

Axiom four requires that betting preferences be independent of the specific consequences that define the bets.

Three further axioms required by Savage, Nondegeneracy, Small Event Continuity, and Uniform Monotonicity are needed only for technical reasons, and need not be specified for interpretations of later work in this thesis.

Savage argues that the applicability of SEU is restricted to what he calls a "small world". A small world is one within which there is always a possibility for the agent to "look before you leap" (Binmore 2009, p. 117; Ross 2014, p. 239). In this world, it is arguably possible for an agent to take account in advance of the implications of all conceivable future information for her subjective utility. Updating information therefore can correct mistakes in her model.

RPT is best understood philosophically as an application of the IS.

RPT is frequently criticised in economics as a form of instrumentalism, echoing the most common objections to the IS among philosophers.[11] This criticism states that modelling choices using RPT is just a way of applying math: utility or preference identified by RPT does not really exist. According to this critique, RPT is an instrument instead of a model of reality.[12] The critique is mainly based on two arguments: first, preferences cannot be revealed by choice alone as RPT suggests; second, RPT fails to support the assumption that behaviour is caused by mental states (Hausman (1992, 2000, 2008); Sen (1973, 1980, 1993, 1997)).

[11] See Ross (2014), p. 69.
[12] For critiques about RPT along these lines, see Hausman (1992, 2000, 2008) and Sen (1973, 1980, 1993, 1997). For a detailed survey of the debates, see Hands (2013).

To refute those arguments, we must adopt externalism plus Dennettian realism about the IS. In fact, the main issue at stake is the reality of the mind. Dennett and Ross maintain that what is real about mind is accessed through observable behaviour by taking the IS to identify real patterns, whereas Hausman et al. maintain that internal mental states, which are unobservable, are constitutive of real mental life.

Given the interpretation of RPT as an application of the IS, and given the defense of the realist interpretation of the IS as opposed to the instrumentalist interpretation of the IS, we can apply the same realist interpretation to RPT. If IS realism stands, then the utility function revealed from an agent's choices, according to RPT, can be regarded as real, although not as the cause of choices. A utility function constructed on the basis of RPT is, rather, a model of the choice themselves (Binmore 2009).

The key point of the alignment between the IS/RPT and Bicchieri's theory of social norms is that, for modelling social norms, there is no appeal to private preferences about how society should be, instead it is the networks of expectations as social facts which condition the agents' behavioural responses, which can be attributed from taking the IS. Norms are theoretically on all fours with any other contingent environmental factors that condition choices.

In the experimental design for this thesis, subjects' choices will be analysed with the assumption that they represent efforts by agents to optimise their expected utility, conditional on their beliefs.

Chapter 4 - Toolbox: How to Investigate Bicchieri's Theory of Social Norms Using a Lab Experiment

Many experimental economists have provided elements of an experimental toolbox for the elicitation of social norms. This chapter will first provide a critical review of some key experimental work by Bicchieri and co-authors on their operationalisation of Bicchieri's account of social norms. Then, I will provide a critical review on the most widely applied experimental toolkit for norm elicitation in the literature, due to Krupka and Weber (2013). I then introduce the toolbox introduced in my thesis, for the investigation of social norms following Bicchieri.

4.1 Critical Review of Bicchieri's Experiments on Social Norms

Various experimental studies by Bicchieri and co-authors on norms have aimed to provide data to evaluate her theory of social norm: see Bicchieri and Xiao 2009; Bicchieri and Chavez 2010, 2013; Bicchieri, Xiao, and Muldoon 2011; Bicchieri and Zhang 2012. There are three main flaws in these experimental designs. First, both empirical and normative expectations are measured by means of modal responses, rather than distributions. For example, in Bicchieri and Xiao (2009, p. 201) subjects were asked "How many dividers[13] in this room do you think answered "Yes" to question (d)? (If your answer is the same as the actual number, you will receive an additional $1)"; in Bicchieri, Xiao and Muldoon (2011) subjects were asked "What is the option that you think most participants chose today? (If your answer is right, you will get one point)". Questions asked in such format do not elicit distributions of beliefs, and so are insensitive to degrees of confidence in beliefs. Second, their experiments fail to apply adequate incentive control. Third, the experimental work by Bicchieri and co-authors employs a simplistic understanding

[13] "Divider" is the term applied in their experimental instructions, which refers to "dictator" or "proposer" in the dictator game.

of belief consistency, reviewed in Chapter 2 section 2.

In this section, I provide critical reviews of three highly cited experimental studies by Bicchieri and co-authors, followed by explanations of the three limitations. The aim is to demonstrate the misalignment of Bicchieri's philosophical work and her empirical work on social norms.

4.1.1 Review of Bicchieri and Co-authors' Empirical Work

Bicchieri and Xiao (2009) (BX) conducted a dictator game (DG), and focused on the role of dictators ("dividers" in their instructions). In a DG, a proposer/divider is asked to divide an endowment between herself and a (typically anonymous) partner. The recipient can, in turn, accept or reject the proposer's offer. The recipient's decision has no influence over the outcome of the game. They provided two types of information to the subjects in the dictator's role: information about what other participants in the dictator role *do* (fair/selfish choice), and what other participants *think ought to be done* (fairness/selfish choice). By offering this information, they intended to influence dictators' empirical and normative expectations to be either in the direction of selfishness or fairness, then compared their choices under the different expectations caused by different information. Their experiment follows the design of Xiao and Houser (2009), with the action space for dictators including splits of $10 into ($9, $1), ($8, $2), ($6, $4), ($5, $5), ($4, $6), ($2, $8), ($1, $9). Both Xiao and Houser (2009) and BX interpret divisions that provide $5 or $4 to receivers as *fair*, and those that provide $2 or $1 to receivers as *selfish*.

Both types of information provided to dictators are summaries based on data drawn from Xiao and Houser (2009).[14] Information about *actual choices* was given

[14] Xiao and Houser (2009) implemented a DG, in order to test whether *ex post* opportunities for emotion expression by the receivers (sending message to dictators after receiving the dictators' divisions) can reduce dictators' profit-maximising decisions. Part of their data on dictators' behaviour is applied in BX as the basis for the message to induce subjects' expectations of fair offers.

to dictators by statements such as "60% of the dividers who participated in a session of this experiment last year *shared* the amount approximately equally (i.e., choose an option that allowed their counterpart to get 40% or more)". Information about normative expectations was given to dictators by statements such as "60% of the dividers who participated in a session of this experiment last year *said* that dividers *should* share the amount approximately equally (i.e., chose an option that allowed their counterpart to get 40% or more)". The message contents varied and were distinguished by 6 treatments: Fair belief (FB), Selfish belief (SB), Fair choice (FC), Selfish choice (SC), FB+SC, and SB+FC.[15] The first four treatments aimed at influencing a single type of expectation in a single direction of either fairness or selfishness, and the last two treatments aimed at influencing empirical and normative expectations in conflicting directions.

In the experiment of BX, each subject played the DG, then answered partially incentivised survey questions for belief elicitation. The DG in this study was a standard one. The belief questions asked about the empirical and normative expectations of dictators and receivers. There were three sets of questions for dictators: the first type of question asked each dictator to choose 1 out of 2 options[16] to answer the question about how many dictators they believed *split* the money approximately equally (i.e., gave the receiver $5 or $4). The second type of question asked about each dictator's first-order and second-order normative responses, specifically, 1) whether they thought dictators *should* split the money approximately equally; and 2) to answer how many[17] dictators they believed answered "yes" to

[15] The message for the FB treatment is "60% of the dividers who participated in a session of this experiment last year said that dividers **should** share the amount approximately equally (i.e., choose option that their counterpart gets 40% or more)"; for the SB treatment the message is "60% of dividers who participated in a session of this experiment last year said that dividers **should** approximately maximize their own earnings (i.e., choose option that their counterpart gets 20% or less)"; for the FC treatment the message is "60% of the dividers who participated in a session of this experiment last year shared the amount approximately equally (i.e., **chose** option their counterpart got 40% or more)"; and for the SC treatment the message is "60% of the dividers who participated in a session of this experiment last year approximately maximized their own earinings (i.e., **chose** option that their counterpart got 20% or less)".

[16] It is not specified whether the choices were given as a proportion or a specific integer.

[17] This is an open question. Again, it is not specified whether the number for the answer is expected to be a proportion or a specific integer.

question 1). The third type of question asked each dictator what they thought their receivers believed they would and should choose.

The belief questions to receivers also consisted of three types. First, what option did they think their counterpart divider *would* choose? If the subject's answer was the same answer that her counterpart wrote on her survey before she knew her final decision, the subject would earn an additional $1. Second, what option did they think their counterpart *should* choose? If a subject's answer was the same answer that her counterpart wrote on her survey before she knew her final decision, the subject would earn an additional $1. The third question elicited a conditional normative expectation which asked: if each receiver's counterpart dividers *knew* that 60% of the *dividers* who participated in a session of this experiment last year *said* that *dividers should* share the amount approximately equally, what option did they think divider *would* choose? The options were splits of $10 into ($9, $1), ($8, $2), ($6, $4), ($5, $5), ($4, $6), ($2, $8), ($1, $9).

The incentivisation mechanism applied in this study worked as follow. Each subject was paid a $5 show-up fee. Each subject earned the amounts gained from the DG task. Expectation elicitations referring to dictators' beliefs about the fair choices paid out $1 if a subject's answer matched the actual number of fair choices. Expectation elicitations referring to dictators' beliefs about receivers' beliefs paid out $1 if a subject's answer matched the receiver's answer. Expectations referring to receivers' beliefs about what their dividers would do paid out $2 if a subject's answer matched their divider's actual decision. And receivers' first-order normative expectations and conditional normative expectations were not incentivised.

BX first calculated and compared the mean *expectations* across the 6 treatments. Their results show that when only one message (either about other dictators' beliefs or choices) is presented, both empirical and normative expectations are affected. For example, in the fair choice (FC) treatment where dictators were only informed about the choices from the previous study, the majority (60%) of dictators

made a fair offer, dictators expected 64% of dictators to make fair offers, and expected 68% of dictators to believe that fair offers *ought to* be made. In contrast, in the selfish choice (SC) treatment, when dictators were only told that a majority of dictators (60%) in a previous session made a selfish offer (i.e., gave $2 or $1), dictators expected that only 37% of dictators *would* make a fair offer, and that just 41% of dictators believed fair offers *ought to* be made as well. Similarly, dictators' EE (Empirical Expectations) and NE (Normative Expectations) of fairness are 18% and 25% in the selfish beliefs (SB) treatment, and both expectations are much higher in the fair beliefs (FB) treatment (60% and 67%, respectively).

The results of comparison of the *expectations* elicited in this study show higher empirical and normative *expectations* of fairness in the FC treatment (64% and 68% respectively) than in the SC treatment (37% and 41%, respectively), and a larger difference between the FB treatment (60% and 67%, respectively) and the SB treatment (18% and 25% respectively). In these four treatments, both normative and empirical *expectations* change in the same direction. This is not the case for the FB+SC and SB+FC treatments. The distinctive finding reported is that in the FB+SC treatment the normative expectations (of fair offers) are significantly higher than the empirical expectations (of fair offers).

Similar aggregate analysis was then applied to the *choice* data. The percentage of fair offers is lower in the SB and SC treatments and much higher in the FC and FB treatments. Specifically, the percentages of fair offers are 33% in the SC and 21% in the SB treatments, while the percentages of fair offers are 52% in the FC treatment and 48% in the FB treatment, respectively. The calculated statistics also show that the percentage of fair offers in the SB+FC treatment is significantly higher than in the SB treatment (45% vs. 21%, Z-test, one-tail p=0.05) but not significantly lower than in the FC treatment (45% vs. 52%, Z-test, one tail p=0.32), and that the percentage of fair offers in the FB+SC treatment is closer to what is observed in the SC than in the FB treatment, although neither of these comparisons yields a

statistically significant difference.

BX then applied *probit* regression with marginal effects evaluated at the means of the independent variables. The independent variables are the *EE* of dictators, *NE* of dictators, and *receiver's EE* of receivers. BX argue that these results show that, comparatively speaking, dictators' empirical expectations of fair offers have a statistically significant and positive effect on the probability that a dictator provides a fair offer (marginal effect of 0.012). In contrast, normative expectations of fair offers have statistically insignificant effects on the probability of compliance with a fairness norm, with a *p*-value of 0.132 and marginal effect of -0.006.

BX conclude that both the *probit* and aggregate analyses provide convergent evidence that empirical expectations about other dictators' behaviours, but not normative expectations about other dictators' beliefs, are significant predictors of dictators' behaviour.

This experiment involves limitations in the extent to which it offers support for Bicchieri's philosophical analysis of social norms. First, the beliefs reported in this experiment are elicited as modes instead of distributions. Asking their subjects "How many dividers in this room do you think split the money approximately equally?", does not elicit beliefs as distributions, and thus provided no information about variations in subjects' degrees of confidence in their expectations.

The second limitation in this study concerns incentivisation. The incentives in their experiment are salient but not dominant. Salience here means that the incentives vary with subjects' performance in the task (for example, their subjects earn an additional $1 if their estimates for the belief questions match the real choices made in the DG). The lack of dominance here means the incentives may not be noticeable enough to motivate subjects to make an effort to assume the belief questions accurately. In their study, the lack of dominance refers two features. First, a reported dictator's belief that matches the choice data only earns $1 for subjects who are American college-students. It is doubtful whether such a small incentive

magnitude is sufficient to motivate subjects to give thoughtful reports. Second, the belief questions in their study asked subjects to predict an *exact* proportion or integer (through choosing 1 out of 2 options) in order to earn payment for beliefs. This further undermines dominance, since subjects' expected earnings are multiplied by the probability of an exactly correct guess, so the incentives are even smaller than $1.[18]

Bicchieri and Chavez (2010) (BC) applied a modified version of the Ultimatum Game (UG) aimed at testing the hypothesis that fair behaviour is context-dependent. A UG is a take-it-or-leave-it game, in which a proposer ("Proposer") is asked to divide an endowment between herself and a (typically anonymous) partner. The recipient ("Responder") can, in turn, accept or reject the proposer's offer. If the Responder accepts the offer, then both the Proposer and the Responder keep the amounts chosen by the Proposer. If the Responder rejects the offer, then both the Proposer and the Responder get nothing.

Their experiments featured 3 treatments in the UG task and 2 treatments in a belief elicitation task.

The UG task in this experiment allows each Proposer to choose a Coin option, apart from the (5,5) and (8,2) options, to send to her Responder. In their experiment each Proposer receives an endowment of $10, so a (5,5) split means proposing $5 for the Proposer and $5 for the Responder; a (8,2) split means proposing $8 for the Proposer and $2 for the Responder. A Coin choice here lets the outcome of a fair coin toss determine the proposal: Heads corresponded to (5,5) and Tails to (8,2).

There are three treatments to the choice space for Proposers in BC's UG task: a *full information* condition, a *private information* condition and *a limited*

[18] Assume someone has a diffuse and completely uninformed belief, and there are K bins, and subjects get paid $M if they answered a question correct. Then the payoff from reporting is $M/K in expectation. In Bicchieri and Xiao (2009), the belief questions gave 2 options, so K=2. If all their subjects' beliefs are diffuse and completely uninformed, then the probability that a subject gets a correct guess is 50%, and the expected payment is $0.5 (given that the incentive for dictators' beliefs is $1). This means that the expected payment is even lower if subjects' beliefs are not diffuse, but are distributions for example.

information condition. In the *full information* condition both Proposers and Responders knew there was the option of a Coin flip, and whether Proposer chose Coin or one of the other strategies. In the *private information* condition the Responders did not know that the Coin strategy was available to the Proposers, but the Proposers knew. In the *limited information* condition both Proposers and Responders knew that the Coin strategy was available, but the Responders would not know whether the split chosen by the Proposers was a deliberate choice by the Proposers or was the consequence of a Coin flip.

Each participant played under all three treatments. The procedure was as follows: all participants randomly drew ID codes labelled A1, A2,….An and B1, B2, …Bn. Then, based on codes they held, participants were separated into two rooms based on the random assignment: Room A were Proposers, and Room B were Responders. All participants then played three treatments of the UG task, each round with a different randomly selected person in the other room. Two out of three games at the end of the study were selected to determine cash payments.

The two main design features of the belief elicitation task in this study are as follows. First, all Responders were asked to report their first-order empirical expectations (1D) and first-order normative expectations (1N). This is referred to as the "non-salient" treatment. The non-salient treatment aimed at assessing whether there was an agreement in Responders' normative expectations, an indicator of (as well as necessary condition for) the existence of a social norm, according to Bicchieri's theory. The first-order empirical expectations are three lines of questions[19] which asked Responders to guess what proportion of Proposers they thought would choose each of the options, i.e., (5,5), (8,2), and Coin (the Coin option was omitted in the private condition as the experiment design requires). Responders' correct forecasts to each line of the question would earn $1. The first-

[19] "Three lines of question" means that each line of question asked about one of the options: either (5,5), or (8,2) or Coin.

order normative expectations are also asked through three lines of questions which asked Responders to mark all actions considered fair. This task was not incentivised. Subjects were allowed to mark none of, or one, or more than one of the options as fair.

The second design feature of their belief elicitation task is referred as the "salient" treatment. The salient treatment was designed to (1) make the fairness norm, elicited from the non-salient treatment, more salient; and (2) to test for an agreement between Responders' normative expectations and Proposers' beliefs about them. This treatment required Responders and Proposers in half of their sessions to answer two extra incentive-based questionnaires. These questions asked participants (in both roles) about their beliefs about the percentage of Responders who marked (5,5), (8,2) and Coin, respectively, as fair options. These are the second-order normative expectations referring to the first-order normative expectations that asked Responders what they believed was fair.

Each subject first answered the questionnaires for the belief elicitation task, with order varying between the *non-salient* versus *salient* treatment. Then all subjects played three rounds of the UG task, with a randomly matched different person in the other role in each round, following the order of *full information* condition, *private information* condition, and *limited information* condition.

BC hypothesised that, if there is a social norm, there should be agreement between Responders' normative expectations and Proposers' beliefs about them. The agreement here refers to an alignment between first-order normative expectations about Responders elicited in the non-salient treatment and second-order normative expectations about Proposers. BC also hypothesised that Proposers' choices in the UG task would depend on the information condition and salience. With respect to the second hypothesis, BC made three directional predictions. First, they predicted that the proportion of Coin choices would be higher in the full information condition than in other two information conditions. The rationale behind this prediction is that

in the private information condition, there are no normative expectations for the Coin option. Second, they predicted that there would be more (8,2) choices in the limited than the other two information conditions because BC think that some Proposers would take advantage of the ambiguity of the choice setting in the limited information condition. Third, they predicted that there would be more (5,5) choices in the private than in the other two information conditions because BC expect that the (5,5) choice is most universally regarded to be fair.

In testing the first hypothesis, BC report two main findings. First, concerning the first-order normative expectations elicited from all Responders in the *non-salient* treatment, almost *all* Responders considered (5,5) to be fair in all information conditions, and a *majority* of them thought that Coin was fair (in full information condition and limited information condition).[20] Table 4.1 shows these results.

Table 4.1: "Normative expectations" of Responders. Each cell contains the proportion of Responders who indicated that the choice was fair in the *non-salient* treatment.

	Choice					
Condition	5,5		8,2		Coin	
Full	96.4%	27/28	14.3%	4/28	64.3%	18/28
Private	96.4%	27/28	17.9%	5/28		
Limited	96.4%	27/28	14.3%	4/28	57.1%	16/28

The second main reported finding is that there is "a remarkable degree of agreement" (p. 169) between Responders' and Proposers' second-order normative beliefs in the *salient* treatment about the normative expectations of Responders in the *non-salient* treatment. In the *full information* condition, the mean of Responders' beliefs about fairness (second-order normative expectation about Responders first-order normative expectation of fairness) were 97.0%, 12.6%, and 63.6%, for the (5,5), (8,2) and Coin options, respectively. The mean of Proposers' beliefs about

[20] This is because that there is no belief questions on the normative expectation of the Coin option in the Private information condition.

fairness (second-order normative expectation about Responders first-order normative expectation of fairness) were 96.6%, 14.9%, and 65.0%, for the (5,5), (8,2) and Coin option, respectively. In the *private information* condition, the mean of Responders' beliefs about fairness were 98.1% and 16.0%, for the (5,5) and (8,2) options, respectively. The mean of Proposers' beliefs were 99.1% and 12.5%, for the (5,5) and (8,2) options, respectively. In the *limited information* condition, the mean of Responders' beliefs about fairness were 96.0%, 10.0%, and 54.4%, for the (5,5), (8,2), and Coin options, respectively, and the mean of Proposers' beliefs were 98.8%, 17.6%, and 49.3%, for (5,5), (8,2), and Coin option, respectively.

In comparing the two main conclusions about the belief data summarised above, BC claim that their data shows that participants' beliefs (second-order normative expectations) from the *salient* treatment about the first-order normative expectations are *in agreement* with the first-order normative expectations themselves from the *non-salient* treatment. This is claimed to provide support for their first hypothesis that there is *agreement*/*consistency* between Responders' first-order normative expectations and the Proposers' second-order expectations about them.

In order to test the second hypothesis, BC applied a multinominal logit model to their choice data, controlling for information and salience. They found small effects of salience, so they averaged across salience conditions in reporting the choice proportions between different information conditions. The results are as follows:

- 37.7% (20/53) of Proposers chose Coin in the full information condition, compared to 11.3% (6/53) in the private information condition and 5.7% (3/53) in the limited information condition. This is argued to be consistent with their first prediction following the second hypothesis, noted earlier.
- More Proposers chose (8,2) in the limited information condition (58.5% [31/53]) than in the full (24.5% [13/53]) or private information (37.7% [20/53]) conditions. This is claimed to be consistent with their second

prediction, noted earlier.

- The highest frequency of (5,5) choices was in the private condition (50.9% [27/53]) compared to the full (37.7% [20/53]) and limited (35.8% [19/53]) information conditions. This is considered to be consistent with their third prediction, noted earlier.

BC conclude that their data support both of their hypotheses. The statistical support for their first hypothesis is declared to be composed of two aspects: consistency between the Responders' first-order normative expectations about fairness, and consistency between the Responders' and Proposers' second-order normative expectations about Responders' first-order normative expectations about fairness. The support for their second hypothesis indicates that the existence of belief consistency (particularly, the second type of consistency) follows from support of their three predictions of their subjects' behaviour. BC further conclude that manipulating the expectations produced shifts in norm-abiding behaviour.

Unfortunately, several limitations in this study undermine these inferences. First, given the important role that second-order normative expectations (2N) play in Bicchieri's philosophical analysis of social norms, BC failed to assess second-order normative expectations. As shown in their Appendix D, Responders were asked to "guess how many (in percentage) Responders (excluding the subject herself) will select each option as a fair option". This refers to beliefs about first-order descriptive expectations. Bicchieri (2017) identifies the first-order descriptive expectation as "what one believes about what others do", while the second-order normative expectation as "what one believes about what others believe I/others should do" (p. 70). In fact, BC never elicited second-order normative expectations as the theoretical analysis suggests they needed to.

Second, their inference concerning belief consistency requires a simplistic understanding of consistency. All the belief reports in this experiment are elicited as modes instead of distributions. BC seem to consider the mode values being similar

as equal to “belief consistency”. This ignores the fact that belief consistency must be measured in the statistical sense, with consideration of belief distributions, instead of beliefs as point estimates (i.e., modal values). The methodology applied in their study with respect to belief elicitation makes it impossible to generate such inferences.

Third, limitations regarding incentivisation are the same problems reviewed previously with respect to BX. The incentives in BC’s experiment are salient but not dominant. And, the belief questions in their study asked subjects to predict an exact proportion in order to earn payment for beliefs.

Xiao and Bicchieri (2010) (XB) conducted a trust game (TG), with a focus on the role of trustees’ perception of the kindness of investors’ actions, in order to explore the effect of “self-serving bias” under conflicting social norms. In a TG two subjects are paired. A trustor (sometimes called investor, such as in the experiment of XB) decides how to divide an endowment between herself and her paired partner, the trustee. The amount the trustor transfers is usually tripled by the experimenter and added to the trustee’s earnings. As a response, the trustee can transfer any amount back to the trustor. There are many variations of the TG conducted in the economics literature. XB make use of the concepts of inequality aversion and reciprocity (hereafter equality and reciprocity) from the economic literature (Fehr & Gächter, 2000a), which are considered to motivate behaviour in society. In contrast to BC (2010), where two norms are available (equality and equity), XB constructed their study so that equality and reciprocity norms are in conflict. XB hypothesise that self-serving bias exists in the TG context such that individuals tend to put greater weight on the norm favoring themselves. In addition, XB propose that a common assumption in economics is to assume individuals are profit maximisers. Hence, there are three theories to compare: profit maximisation, inequality aversion and reciprocity. Because the profit maximisation theory is easy to test (through

calculation),[21] XB hypothesise that back transfers are motivated in part by reciprocity and in part by equality. XB constructed their experiment in a way that equality and reciprocity motives for trustees cannot be reconciled. This design avoids the confounds between the motives of reciprocity and inequality aversion with the motive of profit maximisation.

The experiment has two main tasks: a TG task and a belief elicitation task.

The TG task includes two treatments, a baseline treatment and an asymmetric treatment, differing in terms of the endowments given to their subjects. The baseline treatment is similar to a standard TG game in which both investor and trustee were given 40E$,[22] the investor can transfer one out of two amounts: either 0E$ or 10E$ to the trustee. The amount transferred from the investor to the trustee gets tripled when received by the trustee. So, if the investor transfers 10E$ to her trustee, the trustee gets 30E$. Trustee can send back an amount that is any multiple of 5E$ between 0E$ and 30E$, so the choice space for the trustee is 0E$, 5E$, 10E$, 15E$, 20E$, 25E$, and 30E$. The asymmetric treatment has the same design except the different endowment: 80E$ for investors and 40E$ for trustees. In this treatment it is hypothesised that the two motives, reciprocity and inequality aversion, lead to different decisions comparing with the decisions subjects would make in the baseline treatment. By comparing trustees return behaviour across the two treatments, XB intended to draw different inferences about decision making under the two conflicting norms.

The belief elicitation task asked incentivised belief questions to both investors and trustees about the option of 10E$. The questions to investors were three first-order descriptive and normative expectations. The first type of first-order descriptive

[21] For example, in the ultimatum game, proposers usually propose a non-zero split to their responders and the responders often reject selfish offers, even if this means they would lose the amount transferred from the proposers' non-zero offers.

[22] E$ means experimental currency. Earning in experimental currencies are converted to payment in cash at the end of experiment. The exchange rate between the experimental currency and real currency are determined by the experimenters, and in this experiment the conversion rate is 5E$ = 1 US Dollar.

expectations asked investors ("Actor 1" in their experiment) to anticipate how many Actor 1's they believe transferred 10E$. The second type of first-order descriptive expectations asked investors to answer how much trustees ("Actor 2" in their experiment) they think would return to them, if they (investors) transferred 10E$. The first-order normative expectations asked investors to report how much they think Actor 2s should return to themselves, given they (the investors) transferred 10E$.[23] Subjects would earn an additional $1 if their answer matched the actual number in each question.

The questions to trustees included three second-order descriptive and normative expectations. The first question was a second-order descriptive expectation asking trustees what amount they believed their investors thought they would return. The second question was a second-order normative expectation asking trustees what amount they believed their investors thought they should return. The third question was a second-order descriptive expectation asking what amount they believed other investors would return. Subjects would earn an additional $1 if their answer matched the actual number in each question.

Subjects were randomised to participate in one of the two treatments in the TG task, and each subject played the game only once. The belief elicitation task was after the TG task. Since investors could only transfer 0E$ or 10E$, a transfer of 0E$ means the TG ends (a trustee won't be allowed to transfer anything if her investor transferred 0E$), so the belief questions to investors only asked the investors who transferred 10E$, and the belief questions to the trustees only asked the trustees whose partner investors transferred 10E$ to them. There is a slight difference in the belief task between the investors and trustees: for investors who transferred 10E$, their trustees' back transfer decisions were revealed after they answered the belief surveys; for trustees whose partners transferred 10E$, the

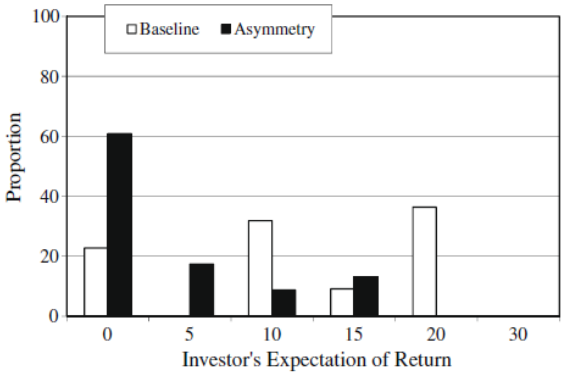
[23] "Actor 1" refers to investors, "Actor 2" refers to trustees . However, as a common practice, in experiments the terms "investor" or "trustee" are replaced by more neutral language in order to avoid misleading subjects' understanding of the experimental tasks.

trustees indicated her back transfer decision before answering the belief questions.

XB reported a similar percentage of investors who transferred 10E$ in both baseline and asymmetry treatments (60.5% and 64.7%, respectively). However, investors' expectations in both treatments differed a great deal. The expected back transfer of the investors who transferred 10E$ in the baseline treatment were 22% (for 0E$), 0% (5E$), 32% (10E$), 8% (15E$), and 38% (20E$), whereas the expected back transfer in the asymmetry treatment were 61% (0E$), 18% (5E$), 9% (10E$), and 32% (15E$).

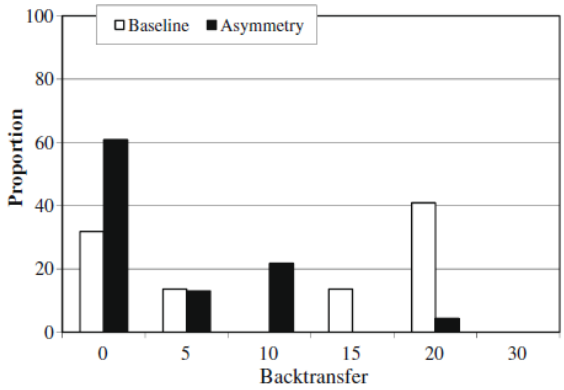
XB claimed that the data of investors' behaviour and expectations have two implications. First, the fact that 2/3 of investors transferred a positive amount is inconsistent with the theory which assumes individuals are profit maximisers since, in a one-shot game, transferring a positive amount to others means the investors suffer from a loss with no expectation of it paying off in future rewards. Second, the fact that 61% of investors who transferred 10E$ expected to return in the asymmetry treatment, a big drop compared with the expectations in the baseline treatment. This is regarded as pointing to consistency with the inequality aversion model instead of reciprocity. The comparison of investors' first order descriptive expectations in the baseline treatment and asymmetry treatment is shown in the Figure 4.1, which is from BC (P. 463).

Figure 4.1: Investors' first-order descriptive expectation of trustees back transfer



Trustees' decisions in the TG task are shown in Figure 4.2, which is from BC (p. 463):

Figure 4.2: Investors' first-order descriptive expectation of trustees back transfer



XB argue that trustees' back transfer behaviour is consistent with their view that back transfer in both treatments are motivated in part by reciprocity, while the inequality aversion motive encourages positive back transfer only in the baseline treatment.

Data on trustees' second-order descriptive and normative expectations about investors are shown in Figure 4.3 and 4.4, which are from BC (p. 464):

Figure 4.3: trustees' second-order descriptive expectation of investors' expected back transfer

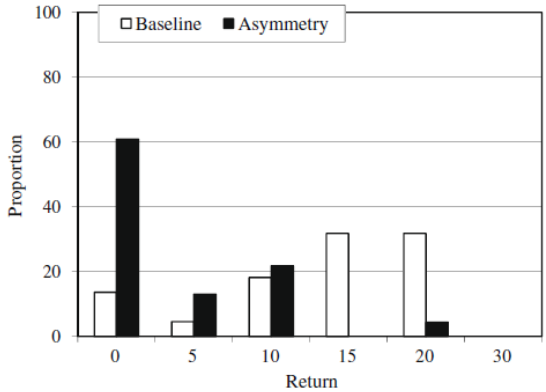
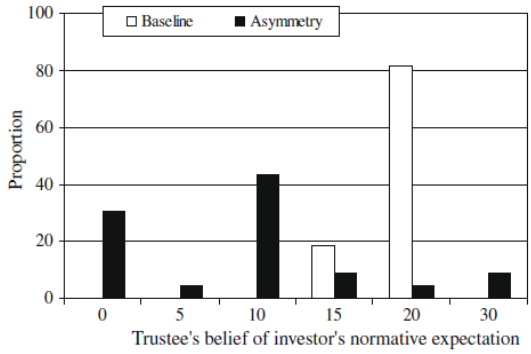


Figure 4.4: trustees' second-order normative expectation of investors' first-order normative expectation



XB conclude that through their experimental study, they obtain observations to support that when two relevant social norms are available, individuals tend to be biased towards the norm that better serves their self-interest. However, methodological limitations cast doubt on the reliability of the data they used to draw this conclusion. First, the belief questions in this experiment are elicited through modal responses rather than distributions. As my previous reviews point out, modal responses can only generate point estimate, which causes insensitivity to subjects' confidence in their beliefs. Second, the experimental design in this study fails to apply incentive-compatibility, which is explained below.

4.1.2 Belief Elicitation and Insensitivity to Agents' Degrees of Confidence

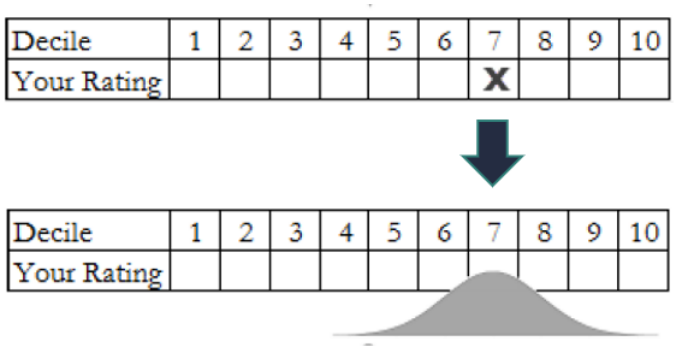
The first limitation in the empirical work by Bicchieri and co-authors is that beliefs are elicited as point estimate instead of distributions. This means that there are no data on agents' degrees of confidence in their beliefs.

In Bicchieri and Xiao (2010), and Bicchieri, Xiao and Muldoon (2011), subjects' beliefs are elicited as subjective probabilities associated with binary events. The experiment in Bicchieri, Xiao and Muldoon (2011) was to determine the normative status of trust. Following Bicchieri (2006), the survey was designed to elicit individuals' expectations about punishment so as to inform the experimenters as to whether trust is a norm. Subjects were firstly described a trust game which was

previously conducted before the belief task. Then they participated in the belief task. In the belief task participants were asked two questions: whether they would impose a fine on either the investor or the trustee, and what their expectations were about what the other participants would choose to do. A 'payoff cut' instantiates the concept of punishment.[24] Actor 1 is the role for trustor, and Actor 2 is the role for trustee. Subjects were asked, for example, "how many participants in today's session do you think chose 'no payoff cut to Actor 1'?" (p. 183), where subjects received one point[25] if their answers were correct. A belief elicited by this method only reflects the mode of the underlying distribution of beliefs, instead of the full distribution.

The concept of a point estimate versus a distribution estimate is illustrated by Merkle and Weber (2011). Figure 4.5 demonstrates this concept, from Merkle and Weber (2011, p. 264).

Figure 4.5: Point Estimate Versus Distribution Estimate



The top panel of Figure 4.5 represents a belief measured by a point estimate. It shows a modal belief, located at the 7th decile of a belief distribution. This example reflects earlier research where participants were asked to judge their sense of humor, and most people placed themselves in the 7th decile, and they place no probability in

[24] A payoff cut in Bicchieri, Xiao and Muldoon (2011) refers to a fine that trustees were asked to impose on their counterpart trustors, if the trustees were not satisfied with the proposals offered by their counterpart trustors. This method was motivated by discoveries in cross-cultural experimental studies which suggest that many diverse cultures are willing to incur cost to punish counternormative behaviour (Henrich et al. 2005, 2006).
[25] According to the Questionnaire in Bicchieri, Xiao and Muldoon (2011), each point subjects earn is worth $3 (p. 183).

other deciles. In BX (2009), subjects were asked to report their beliefs as a point estimate: "how many dividers in this room do you think split the money approximately equally?". An answer to this type of question is called the *modal value* in statistics.

In comparison, the bottom panel in Figure 4.5 represents a belief in the form of a distribution. It shows a person's belief as actually assigning positive weights across decile 4 to 10. This distribution is symmetric. The highest weights are actually assigned to decile 7, followed by decile 4 and 8 (which have same weight). Decile 3 and 9 have less weight than the amount that decile 4 or 8 have, and the least weights are equally in decile 4 and 10.

The problem with eliciting just a point estimate is that it ignores the confidence of subjects by only taking the mode of a distribution instead of the distribution itself. The dispersion of the actual distribution in the bottom panel presents the agent's confidence level in her belief regarding each possible outcome. This dispersion is called the *standard deviation* of the belief distribution. Bias in a belief report generated from ignoring belief confidence is demonstrated in Figure 4.6, from Harrison et. al. (2021).

Figure 4.6. Bias and Confidence of Subjective Belief Distributions

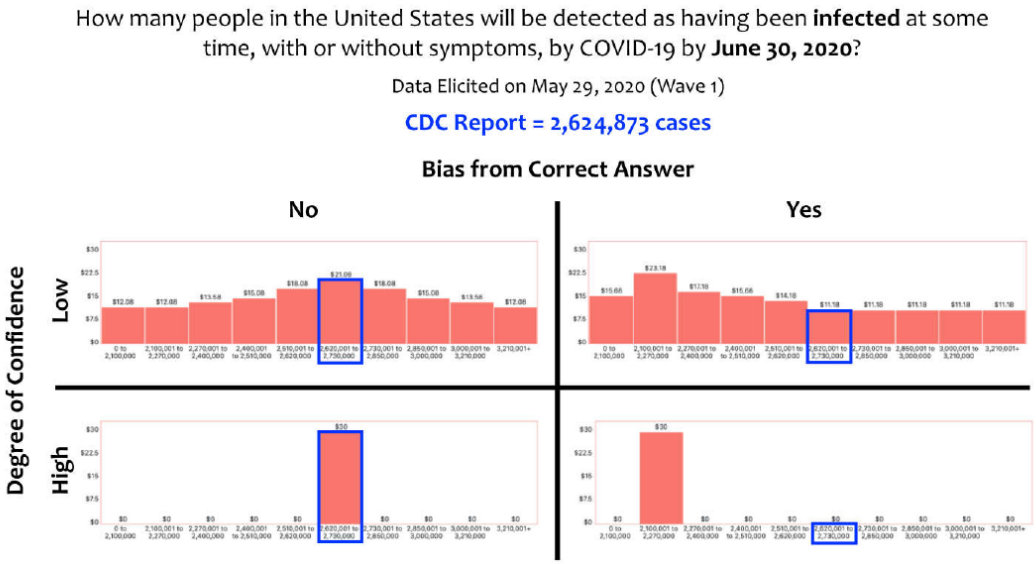


Figure 4.6 demonstrates how the concept of bias interacts with confidence in belief distributions. In statistics, bias usually refers to a point estimate, such as the mean or the mode. On the left side of Figure 4.6, we can see two belief distributions: the top one presents low confidence and the bottom presents high confidence. Both belief distributions are unbiased (the modal value is the same as the correct answer). On the contrary, on the right side of Figure 4.6 we can see that though both belief distributions are biased (the modal value deviates from the correct answer).

4.1.3 Belief Consistency

The second limitation in the empirical work by Bicchieri and co-authors is that they seem to apply an overly simplistic understanding of belief consistency. Bicchieri and co-authors seem to argue that their empirical results support the hypotheses with respect to both types of belief consistency, i.e., *within-individual* belief consistency and *across-individual* belief consistency. However, I argue, for two reasons, that the inference does not stand. The first reason concerns the methodological limitations reviewed in chapter 2.2.1 and 2.2.2. Second, their inferences fail to test belief consistency in the manner of belief distributions.

Bicchieri and co-authors seem to believe their experimental data support their inferences on both types of belief consistency automatically. For example, BX (2009) argue that their experimental findings in the DG provides evidence that when empirical and normative expectations are in line with each other, the choices will follow a social norm (p. 202). This refers to the first type of belief consistency, *within-individual consistency*.

Bicchieri and co-authors also argue that their data shows the existence of an agreement of normative expectations at a social level (between individuals), which is the second type of belief consistency, *across-individual consistency*. For example, Bicchieri and Chavez (2010) argue that their data from a DG implies an agreement of the expectations of the subjects in the Responder role, with beliefs about them in

the Proposer role. That is, according to them, an existence of mutual consistency of the second-order normative expectations *between* Proposers and Responders. Bicchieri, Dimant and Sonderegger (2019) also claimed that they have successfully tested for across-individual consistency of 2N beliefs.

This methodological limitation regarding belief elicitation blocks accurate inferences of the existence of belief consistency in their empirical work. As reviewed in section 4.1.2, in all their empirical work belief distributions are generated as point estimate, instead of estimate of distributions. This causes failures to adequately assess belief consistency with respect to bias and confidence. As a result, the existence of belief consistency claimed in their empirical study, at best, can be understood in a statistical sense as consistency between modal values of the beliefs in comparison.

Hypothesis testing on belief consistency following Bicchieri's philosophical analysis, however, must take the "whole body" distribution into consideration. Belief distributions allow us to evaluate not only bias of the point estimates, but also dispersion of the distribution. Belief dispersions at the level of individual are naturally understood as confidence measures. In statistics, most tests of the consistency of beliefs are only tests for bias at a point value, without considering the entire distribution of the beliefs. I argue that the understanding of belief consistency reflected from the work by Bicchieri and co-authors reflects an overly simplistic view of belief consistency, because their work ignores the statistical concept of belief consistency. In order to test the hypotheses of belief consistency, methods of statistical inference regarding the "whole body" distribution comparison must be applied.

4.1.4 Lack of Salient and Dominant Incentives

One further limitation in Bicchieri's experiments on social norms is the control of economic incentives, the first important aspect of which is **salience**. Salience

refers to rewards that vary with subjects' performance in the task. If incentives aren't salient, the experiments cannot be incentive compatible, and can therefore generate hypothetical bias. For example, in Bicchieri, Xiao and Muldoon (2011), subjects are paid only with a task-fee for finishing just the questionnaire, and there is no variation of the incentives with the subjects' performance in the task.

The second limitation regarding incentivisation in Bicchieri and co-authors' empirical studies is its lack of **dominance**. Dominance means that the rewards, apart from being salient, have to be noticeable to the subjects. Smith and Walker (1993) show that subjects expend more *cognitive effort* with increased incentives. This partially refers to the size of payment in experimental studies. However, in Bicchieri and Xiao (2010) subjects in an investor's role are incentivised with only an extra $1 if the reported empirical belief matches the amount a trustee returned (p. 460); in Bicchieri and Chavez (2010), subjects are rewarded only $1 for a correct prediction over binary questions about an UG task (p. 177). And in Bicchieri, Dimant and Sonderegger (2019) a correct answer is rewarded with payoff of only $0.25. It is doubtful that such a small reward would be *noticeable* enough to act as an effective incentive in generating accurate belief reports.

4.2 An Alternative Experimental Procedure to Elicit Norms

Krupka and Weber (2013) (KW) introduced a methodology for norm identification which has gained popularity in the literature of experimental economics. In this section I will review their method.

KW's study is built on the assumption that decision makers' utility is based on the money they obtain and on the degree to which their actions comply with social norms, in the form of taking actions generally viewed as socially appropriate and avoiding those viewed as socially inappropriate. Therefore, the utility function in their study is defined as follows:

$$U(a_k) = V\big(\pi(a_k)\big) + \gamma N(a_k)$$

In their modelling social norms are regarded as collective perceptions regarding social appropriateness, such that a social norm "$N(a_k) \in [-1,1]$, is an empirically measurable collective judgment that assigns to each action a degree of appropriateness or inappropriateness" (Krupka and Weber 2013, p. 499). The formal definition in this model is as follows: 1) A = $\{a_1, \cdots, a_k\}$ represents a set of *K* actions available to a decision maker; 2) a_k represents an action chosen by an agent; 3) the function $N(\cdot)$ represents an agent's norm adherence. KW assumes that if for an action, a_k, there is a collective recognition that the action constitutes "appropriateness", hence $N(a_k) > 0$; and if an action, a_k, there is a collective recognition that the action constitutes "inappropriateness", hence $N(a_k) < 0$.

The function V $(\cdot)$ represents the value the individual places on the monetary payoff, so function $\pi(a_k)$ is assumed to be increasing. The parameter γ for the norm adherence function $N(\cdot)$ represents the degree to which an individual cares about norm adherence, therefore, as γ increases an individual's utility will increase from selecting actions which are more 'socially appropriate' compared to her utility from not selecting such actions. Hence, $\gamma \geq 0$. The parameter $\gamma = 0$ defines when an individual is unconcerned about the social norm, and only cares about her monetary payoff.

A closer look at this definition reveals a disagreement about the deviation of social norms which will be later seen as a misalignment between theoretical work and experimental studies on the nature of social norms. The definition of social norms used by KW is as follows (p. 499):

> "Following the literature, we define (injunctive) social norms as collective perceptions, among members of a population, regarding the appropriateness of different behaviours. They are things that people in the population *jointly recognize one should* or should not do, and people who belong to the population *expect others to be aware* of and understand this agreement".

Following Bicchieri's language, we can see the definition adopted by KW seems to include both *1N first-order normative beliefs* (that people in a population jointly recognise things one should or should not do) and *2N second-order normative*

beliefs (that people expect others to be aware of the agreement). However, in their experimental work, this distinction seems to disappear.

Given the formal definition of norm appropriateness summarised in their utility function, KW conducted an experimental study using a DG to show how to identify social norms, and to demonstrate how the theory of norm appropriateness can predict behavioural changes across different treatments of the DG. In addition, KW compared their experimental results with other previous studies,[26] to demonstrate that different contextual features of a choice environment can lead to varying social norms in play and various behaviours across different contexts.

The experiments of KW focused on two objectives. Experiment 1 focused on elicitation of social norms by asking subjects to rate the degree of social appropriateness towards each option varying across different contexts. Experiment 2, which is conducted by subjects separated from the ones participated experiment 1, asked subjects to play the DG game across the different contexts in experiment 1. The results from the norm rating reported in experiment 1 are applied by KW to predict how behaviour in experiment 2 would differ between the two environments. The predicted effects were tested by data collected from experiment 2.

KW designed the choice environment of the DG in their experiment 1 in two variations: a standard DG, and a "bully" DG. In a standard DG, the participant in the role as a "dictator" initially receives $10 while her randomly matched anonymous partner receives $0. The dictator must decide how much, between $0 and $10, in one-dollar increments, to give to her partner. In a "bully" DG, the dictator and her randomly matched anonymous partner both receive $5. And the dictator can decide to *give* or *take* any amount between $0 and $5, to or from her randomly matched anonymous partner, again in one-dollar increments.

The key "trick" in the design of experiment 1 is that both choice environments offered the decision maker exactly the same eleven choices over final wealth

[26] See Dana, Weber and Kuang (2007), List (2007), and Lazear, Malmendier and Weber (2012).

allocations ranging from ($10, $0) to ($0, $10), but varied in the actions (*give* or *take*) required to obtain those dollar allocations. For example, for the endowment of $10, in the context of standard DG any outcome other than ($10, $0) requires subject "giving" money to the other person, whereas in the "bully" variation outcomes other than ($5, $5) subjects are "taking" money from the other person. In general, KW conjectured that the action of "taking" would generally be considered less socially appropriate than the action of "giving", even that both actions produce same outcomes.

Subjects in experiment 1 faced six situations, and they were asked to rate the social appropriateness of each action choice available in each situation. This is regarded as a task to elicit a norm by KW. There are four scales of social appropriateness: 'very socially inappropriate', 'somewhat socially inappropriate', 'somewhat socially appropriate' and 'very socially appropriate'.

All the situations in experiment 1 are hypothetical. The incentivisation for experiment 1 works as follows: before subjects reported their ratings, they were told that one of the situations for which they rated would be selected at random for payment after the task ends. Among all the selected ratings over action choices, only ratings that match the modal response of the session were incentivised (with an additional payment of $5 in experimental sessions conducted in Pittsburgh, and $10 in experimental sessions conducted in Michigan).

Experiment 2 is a task in which subjects, separated from the ones in experiment 1, were asked to play one of the two variations of the DG, either a standard DG or a "bully" DG. Subjects in this task were incentivised with a $2 show-up fee in addition to any money from the allocation choices made in the DG variants they were randomly assigned to.

Based on the result from Experiment 1, KW made two behavioural predictions. The first prediction states that "more agents will select the action producing the equal-split ($5, $5) allocation in the "bully" environment than in the standard

environment" (p. 507). The second prediction in KW states that "conditional on not selecting the action producing the equal-split ($5, $5) allocation, more agents will select the action producing the payoff-maximizing ($10, $0) allocation in the bully environment than in the standard environment" (p. 507).

KW claim that their two predictions were all supported by the behaviour data they received from experiment 2. KW argue that experiment 2 demonstrates that behaviour changes significantly across the two choice environments (the standard DG and the "bully" DG), although the realised payoffs across the two choice environments were almost identical. Therefore, KW conclude the behaviour data from their experiment 2 confirms that differences in behaviour (observed in experiment 2) are accounted for by differences in the social appropriateness (data from experiment 1). The data from experiment 1 are regarded as elicited social norms in their study.

The methodology for norm elicitation developed by KW has gained some popularity in the economic literature. For example, Gächter, Nosenzo and Sefton (2013) follow the KW method to model peer effects in a three-person gift exchange experiment. Burks and Krupka (2012) apply the KW method in a real firm to model the norms on-the-job behaviour among financial advisers and their supervisors. Schmidt (2019) also extends KW method to measure not only normative norms but also descriptive norms. However, Schmidt (2019) differs from KW in terms of two experimental design features: a mini-DG instead of standard DG, and a within-subject design for belief tasks instead of between-subject design. Schmidt (2019) argues that their study corroborates that KW's approach is a valid tool for norm elicitation at the individual level, and that the individuals' coordination choices in both injunctive norms and descriptive norms are strongly related to their actual behaviour.

Following the practice by KW and Burks and Krupka (2012), Schram and Charness (2015) (SC) designed a laboratory experiment by examining dictator

choices to see to what extend a shared understanding of peers' advice would affect people's public behaviours. SC hypothesised that peer effects are manifested by a phenomenon that employee's effort is sensitive to the efforts of other employees in the context of the three-person gift exchange game in their experiment. Their study varied between the facts of 1) whether there exist normative expectations, and 2) whether choices are made public. Findings from SC's work show that normalised standard error is lower when there is advice which are the shared understanding of normative expectations. SC's (2015) study took one step further than KW's work by introducing the role of second-order normative belief into the belief elicitations.

Though the KW method of norm elicitation has gained popularity in the literature, it has some limitations. The first issue is that it solely focuses on *1N first-order normative beliefs*. Fehr and Schurtenberger (2018), for example, applied the KW norm elicitation task by asking their subjects "how many tokens should each group member contribute to the project?". When questions are asked in this form, it is impossible to determine whether a respondent's belief is a predictive one or a normative one. Their experiment is designed to distinguish between a punishment treatment and a non-punishment treatment. It makes sense that subjects in the non-punishment treatment might interpret the word "should" as a descriptive matter, instead of a normative matter.

Second, the belief elicitation in this study is point estimates (the modal value), not an estimate with a distribution, which causes the data to be incomplete. In experiment 1 in KW's study, subjects were incentivised if her response to a randomly selected question is the same as the most common response provided in the same session.

For the reasons given above, my thesis will not employ the norm elicitation methodology developed by KW. I propose experimental protocols which can remedy the problems mentioned, and most importantly my experimental design is an empirical operationalisation of Bicchieri's philosophical analysis of social norms,

which differs from the theoretical background of KW's empirical work. The most important difference is the emphasis on all four types of beliefs following Bicchieri's philosophical analysis of social norms. Another modification over KW's experimental work is to suggest an elicitation of distributions applying quadratic scoring rule, instead of eliciting a point estimate.

4.3 Methodology

As analysed in Chapter 1, the philosophical framework of social norm adopted in this thesis is the one proposed by Bicchieri. It suggests a social structure account of modelling social norms and provides possibilities for it being operationalised in the lab and in the wild. This chapter introduces the toolbox which I suggest can better examine the operationalisability of Bicchieri's philosophical analysis of social norm in the lab experiment. I will argue for the use of best practice experimental methodology.

4.3.1 Quadratic Scoring Rule for Subjective Belief Elicitation

4.3.1.1 Subjective Probability for Subjective Belief

Individuals may have different degrees of belief or judgments due to different backgrounds in their knowledge or experiences. In statistics, subjective belief refers to the degree of belief as a quantified judgment (Winkler 1972, p. 16-18). It is represented mathematically by subjective probability. Subjective probability enables one to explain the distribution of the belief towards different events. The concept of subjective belief was formalised by Savage (1972).

Operationally, subjective probabilities are defined as those probabilities that causally lead to an agent to choose some prospects over others when the outcomes of those prospects depend on events that are not yet actualised. A simple prospect here might be a bet or whether one specific event will occur. Since the events are not yet actualised, and there are several possible events and many possible risky bets

about these events, belief elicitation naturally involves risk. In general, we need to elicit or control for risk preference in order to infer belief.

4.3.1.2 Scoring Rule for Binary Events versus Continuous Events

Scoring rules, betting tasks and simulated auctions are the most popular methods for eliciting subjective probabilities for individuals, and for the purpose of studying inferences about subjective beliefs. In economic terms, scoring rules are functions mapping a subject's reported beliefs about a random variable and the ex-post realisation of that random variable into a payoff for the subject. The most popular scoring rule is the quadratic scoring rule (QSR), first introduced by Brier (1950).

The difference between binary events and continuous events reflects a person's belief being about a binary or a continuous event. Borrowing the example given by Harrison and Ross (2016), a question "A typical American male will live 70 or more years. True or false?" is a binary event. However, "How long will the typical American male live?" is its corresponding continuous event. The theory of belief elicitation for continuous events also applies to non-binary, discrete events. For example, the question, "A typical American male will live to be between 0 and 50, 51 and 65, or more than 65. How likely are each of these age categories?" defines 3 events. Beliefs over these 3 events define, in statistics, a probability *mass* function. Beliefs over the continuous event define, in statistics, a probability *density* function.

The framework of scoring rule for binary events can be formally viewed as a trading game between the agent and the reporter, following Savage (1972). The agent gets paid $X if the outcome of the event occurs, the agent gets paid $Y if the outcome of the event does not occur. The payment amount from the agent's report in the binary event is determined by two positive parameters α and β which define the QSR: α determines a reward and β determines a penalty. Assume an event has two possible outcomes A and B which are complements (so if A does not occur, B must

occur), and let θ be the reported probability for A, and let Θ be the true binary valued outcome for outcome A. This means if A occurs Θ=1, and if A doesn't occur Θ=0. Andersen et al. (2014) show that the QSR score or payment penalises the subject by the squared deviation of the report from the true binary valued outcome Θ (p. 212).

For a risk neutral agent, under the assumption of SEU,[27] her report is her belief. This is because under SEU, the Subjective Expected Value of a risk prospect equals to the certainty equivalent of that prospect, therefore the risk premium is zero. This means the reported belief distribution for a risk neutral agent is her latent belief distribution. However, it is well-known that the payment to the belief reports over binary events in QSR can be affected by the risk attitude of an agent.[28] For a risk averse agent, the report will be different from the report of a risk neutral agent. Risk averse agents tend to report their beliefs close to ½ for binary events, with varying degrees of distortions of reports from beliefs caused by different degrees of risk aversion.

In the context of social norms, beliefs are better modelled as probability distributions with continuous events instead of binary event. For example, if we ask a Chinese person to predict whether a young Chinese couple in his social circle will spend all their savings and even bear debts to host a luxurious wedding for the sake of social image, this agent may not be able to report his belief simply as yes/no, but he may refer to his knowledge on the young couple and evaluate the possibility of or say "to what degree" the young couple may follow the norm or not. Scoring Rules for continuous events are then a better model for eliciting such type of beliefs.

Matheson and Winkler (1976) developed the QSR for eliciting people's subjective beliefs for continuous distributions. The scoring rule can induce truthful reports if the agent is risk neutral. Each report in the scoring rule induces a lottery,

[27] SEU refers to the theory of Subjective Expected Utility developed by Savage (1972).
[28] See Winkler and Murphy (1970), Savage (1972) and Kadane and Winkler (1988).

and the risk attitudes affect the inference of the subjective belief reports. The subjects are rewarded for the accuracy of their beliefs, and the definition of the penalty can be severe if the accuracy misses the true answer.

Assume K events, which means the possible response to the question being asked is partitioned into K intervals, and that each agent's report over the K intervals adds up to 100 tokens. As summarised in Harrison et al. (2017) the QSR payment score works as follows: if k is the interval the actual value lies in, the reward in the score is a doubling of the report allocated to the true interval k, and the penalty in the score depends in a "quadratic manner" on the distribution of the reports over K intervals. For all SEU agents, in order to gain the highest reward, she would never allocate any tokens to a bin if her belief to that interval is zero; she will only allocate different levels of tokens to 2 bins if her belief differs for the 2 intervals (hence, if her belief is the same for the 2 intervals, she would allocate the same number of tokens to each bin of the 2 bins). With risk aversion, the agent would behave as if "flattening" the report for those bins she has positive belief of occurring, so as to avoid risk by reducing the variability of utility over different possible outcomes.

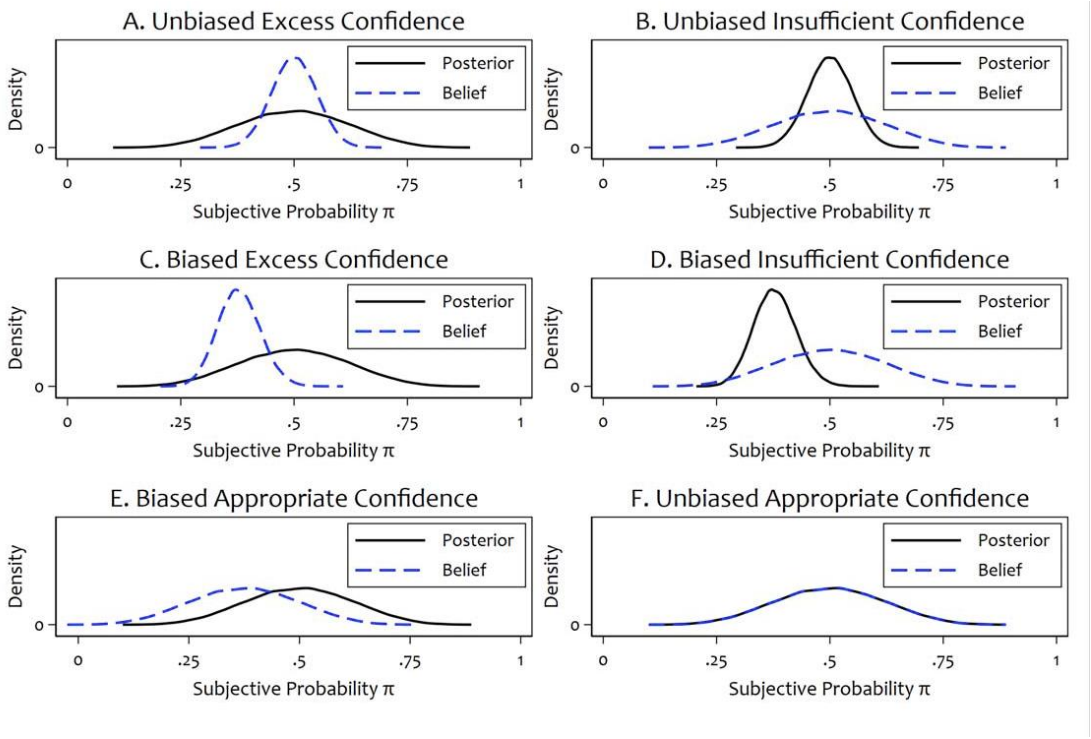
Harrison et al. (2017) ask each subject to conduct a task of allocating 100 tokens by sliding a bar across each of 10 bins that range across the possible realisation values in order to discover their beliefs. We apply a QSR to elicit subjects' subjective beliefs. We use 4-bin belief questions to elicit discrete distributions and 10-bin belief questions to elicit subjects' discretised version of continuous distributions. Each bin is defined by an event if discrete (e.g. 3 out of 4 people) in the 4-bin questions, or by an interval within which the subjective belief might lie if real-valued (e.g., 0% to 9% of all tokens) in the 10-bin questions. The first-order belief questions are 4-bin questions, as we ask for beliefs about how many out of 3 randomly selected people would or should choose to accept or reject a proposed split by the Proposer. The 4 bin options are "nobody", "1 out of 3 people", "2 out of 3 people" and "all people". The second-order belief questions elicit beliefs about the

percentage of tokens allocated by everyone else in the task when reporting their first-order beliefs with respect to the 4 bin options. These second-order beliefs are 10-bin questions, where the intervals of the 10 bins are “0% to 9% of all tokens”, “10% to 19% of all tokens”, “20%-29% of all tokens”,… to “90%-100% of all tokens”. The application of the QSR with the interface developed by Harrison et al. (2017) allows experimenters to study individuals’ belief distributions without burdening the analysis with formulae that most would not understand.

4.3.1.3 Confidence in Subjective Belief Distribution

When we say that someone displays overconfidence, we mean to describe that the person systematically overrates their subjective confidence in their judgements compared to the “objective accuracy” of those judgements, where “objective accuracy” refers to the appropriate Bayesian posterior distributions. Figure 4.7, from Harrison and Swarthout (2021), illustrates the concept of confidence in relation to the concept of bias. In panel A, the subject exhibits unbiased belief from the true answer (with respect to its mean) but presents overconfidence in her belief; panel B shows an unbiased belief with insufficient confidence; panel C demonstrates biased belief with overconfidence; panel D displays biased belief with insufficient confidence; panel E exhibits biased belief with appropriate confidence, and panel F demonstrates unbiased belief with appropriate confidence.

Figure 4.7: Confidence in Subjective Belief Distribution



Harrison and Swarthout (2021) note three distinct definitions of overconfidence in the literature. The first definition is **overestimation** on one's absolute ability to perform a task. For example, one may report a belief of a higher score than one can actually get in a math task. The second definition is **overplacement** of one's performance relative to others. Overplacement is usually known as the "better than average effect", where everyone is relatively better than average. The third definition considers **overprecision** about the certainty in the accuracy of one's belief. Overconfidence in the sense of overprecision refers to the belief expressing unwarranted certainty in its precision.

Overprecision as one of the hypotheses about overconfidence is particularly important to the study of belief elicitation regarding social norms. The need to evaluate the hypothesis about overprecision is explained by Moore and Healy (2008), Merkle and Weber (2011) and Benoît and Dubra (2011). Harrison and Swarthout (2021) argue that it is not clear how inferences about statistically significant overestimation or overplacement can be made without knowing how much precision individuals have about certain beliefs. They point out that because beliefs about

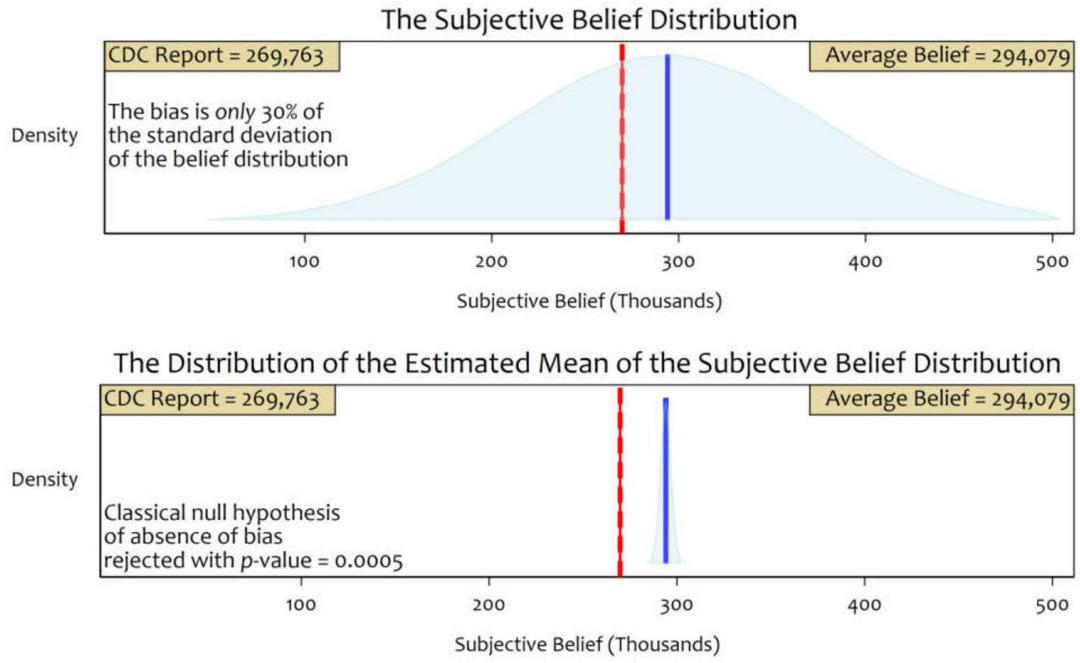
overestimation and overplacement are elicited in the current literature as point estimates, it downplays the importance of sufficient imprecision in beliefs.

The main issue here is that statistical tests about bias of estimates of distributions involve standard *errors*, however confidence about belief distribution involves standard *deviations* of that distribution. For each statistic of the belief distribution the econometric model generates a point estimate and a standard error of that point estimate. One statistic for the normal distribution, assumed here to characterise beliefs, is the mean. So the mean has a point estimate and a standard error. The other statistic for the normal distribution is the standard deviation, so it also has a point estimate and a standard error.

This problem of confusing *standard deviation* with *standard error* causes confusion between tests of the significance of bias of a subjective belief distribution with tests of the bias of an estimate of the mean of the belief distribution.

Confidence is a concept of "imprecision" from statistical estimation. It is most commonly captured by the variance of beliefs about their mean, assuming for now a normal distribution. It could also be regarded as reflecting the variability of beliefs when the beliefs aren't normal distributions. This problem is assessed by Harrison and Swarthout (2021), and it is sharply illustrated by the problem of "precise estimates of imprecise beliefs" (p. 18). Figure 4.8, from Harrison et al. (2021), displays this distinction:

Figure 4.8: Confusing Bias of a Belief Distribution with Bias of an Estimated Statistic of the Belief Distribution



The top panel shows the belief distribution where the bias of the standard deviation in the reported belief distribution is 30% of the estimated standard deviation. The image shows a considerable imprecision of the reported belief. However, the bottom panel shows what causes the problem of precise estimates of imprecise beliefs. The bottom panel is a distribution of the estimated mean of the subjective belief distribution. It applies the classical statistical test and shows that the classical null hypothesis of absence of bias being rejected with *p*-value of 0.0005. The distinction between the top panel and the bottom panel exposes the inferential problem.

In my thesis, I suggest beliefs about social norms should be elicited as distributions and I tend to think of confidence as reflecting the variability of beliefs instead of only paying attention to just the weighted average or mode of beliefs.

4.3.2 Risk Preferences and Subjective Belief Distribution

The use of QSR to elicit beliefs involves giving the subject a choice over one of many possible risky lotteries. When the QSR is applied to a binary event, and the

subject has 100 tokens, there are 101 possible choices. When the subject confirms a choice, she has selected 1 of these 101 risky alternatives. In order to infer her subjective probability over the two events, we need to know her risk preference.

Andersen et al. (2014) shows that risk attitudes will affect the incentive to report one's subjective probability "truthfully" in the QSR. The results from Anderson et al. (2014) show that one has to be sensitive to the risk attitudes of subjects before drawing inferences about subjective probabilities from responses to scoring rules. Harrison et al. (2017) show that a subjective probability distribution can deviate from the reported distribution under QSR applied to non-binary events, and that risk aversion theoretically causes the individual to report a "flattened" version of their true distribution.

An alternative way to account for risk preferences is to apply the Binary Lottery Procedure (BLP) to induce risk neutral behaviour. The scoring rule is predicted to directly induce truthful belief reports only if the agent is risk neutral. The BLP is one methodology to induce risk neutrality. It allows for eliciting subjective probabilities without correcting for risk attitudes. Smith (1961) was the first to suggest the solution of a "probability currency", which allows the inferences of the subjective probabilities over some binary event from subjects' choices over the bets, without necessarily knowing about their utility functions. Roth and Malouf (1979) employed this procedure in many experiments and assumed that we can evaluate the expected utility of two bargaining agents with this device.

Harrison, Martínez-Correa and Swarthout (2013, 2014) and Harrison, et al. (2015) evaluated the BLP in laboratory experiment for inducing risk neutral behaviour in a belief elicitation task. One big advantage of applying BLP in belief elicitation task is to allow the data from the reported belief distributions to be directly applied as subjective beliefs without further need to recover beliefs from observed reports.

Harrison et al. (2013) proposes a simple payment procedure in the lab to implement BLP. In their settings, subjects are asked to make a single choice over a pair of lotteries defined over money and objective probabilities. However, instead of earning money when one of their choices is played out, subjects earn points. These points convert into increased objective probability of winning another binary lottery. Through observed choice patterns, they found that the lottery procedure induces risk neutrality robustly when subjects are given one task.

In practice, the BLP is increasingly applied with the QSR for belief elicitation.[29] In the belief elicitation task, subjects earn points. The payments for the belief tasks are determined by a binary lottery with a larger payoff *M* and a smaller payoff *m*, where more points earned in the QSR translate into a higher probability of receiving *M*. This means, instead of being paid money from their belief elicitation, subjects will eventually be paid by either the large prize *M* or a much small payoff *m*. After a subject finishes their belief task applying QSR, there will be draw of two random numbers between 1 and 100 to determine the outcome of the QSR. The first random number determines the number of points subjects earn and the second random number determines whether the subject wins the high or the low prize according to the points earned.

The methodology of using the QSR with the BLP has been applied in various studies. Harrison, Martínez-Correa and Swarthout (2013, 2014) and Harrison, et al. (2015) evaluated the BLP in laboratory experiment for inducing risk neutral behaviour in a belief elicitation task applying QSR. They found evidence consistent with the theory that by applying BLP risk attitudes have a surprisingly small role in distorting reports from true belief distribution. Hossain and Okui (2013) applied the same method and showed similar results that the reported beliefs are closers to the true probability than the reported beliefs seen under standard QSR when one assumes that reports are beliefs.

[29] The “binarized scoring rule” of Hossain and Okui (2013) is equivalent to the QSR with BLP.

In summary, I argue that risk attitudes have to be elicited and jointly estimated with belief reports in order to recover beliefs *or* that risk attitudes have to be neutralised by experimental design so that the subjects behave as if they are risk neutral for that belief elicitation task. The point is, contrary to Bicchieri and co-authors, one cannot just assume risk preferences away: one has to either jointly estimate *ex post* the data collection for risk preferences and belief reports, or design the experiment *ex ante* to risk-neutralise subjects.

4.3.3 Importance of Incentivisation

In experiments on belief elicitation through survey questions, the first fundamental distinction is whether the surveys are **hypothetical** or financially[30] **incentivised**. Hypothetical choices mean that preferences or beliefs revealed by the choices have no consequences for the decision maker, whereas incentivised choices refer to preferences and beliefs revealed when there are *real* consequences for the decision maker of alternative choices.

The main reason for drawing a clear distinction between hypothetical and incentivised choices is concern about hypothetical bias. Hypothetical bias occurs "whenever there is a difference between the choices made when the subjects face real consequences from their actions compared with the choices made where they face no real consequences from their actions" (Harrison 2014, p. 238). In some choice modelling contexts, however, it is not possible to avoid hypothetical bias. Therefore, it must be mitigated if possible.

Harrison (2014) suggests two ways to mitigate hypothetical bias: by means of instrument calibration *before* the survey or by means of statistical calibration *after* the survey. Instrument calibration focuses on how the hypothetical questions are posed. For example, Blumenschein, Johannesson and Yakoyama (2001) suggested

[30] Incentives could also be in the form of gift card, for example. However, in this thesis, the financial incentives solely refers to monetary incentives.

that survey questions be asked in a way that encourages recognition of some uncertainty in the subject's understanding of what a "hypothetical yes" means. And Cummings et al. (1997), Cummings and Taylor (1998) and List (2001) suggested that the survey questions themselves could encourage subjects to avoid hypothetical bias by telling them about the issue of hypothetical bias. The evidence supporting these procedures is mixed. The approach of statistical calibration was developed by Blackburn, Harrison and Rutström (1994). Statistical bias functions focus on applying some bias functions which attempt to predict the response the subjects would have given if facing incentives after observing what they actually responded when not facing incentives. The usefulness of statistical estimation rests upon the assumption that hypothetical choices are informative. The evidence for statistical calibration is positive so far when applied to valuation of common private goods.

The definition of incentivisation includes three concepts.

The first concept of incentivisation is **salience**. Salience means that the rewards must vary with the performance of the subject in the choice task. That is, the rewards should connect to the *message* subjects send. In the case of belief elicitation, for example, salience means that the token allocations subjects make affect their payoffs. In one scoring rule task for belief elicitation, subjects allocate tokens by moving the slider between different possible outcomes. The fact that the behaviour of a subject of moving the slider changes her payoff in some way means that the reward in the task is salient. Of course, incentives can be non-salient. Non-salient rewards refer to rewards that do not vary with the performance of the subject in the task. For example, if a subject gets paid $20 to complete a task, no matter what answer the subject chooses in the task, then this is a non-salient reward. Smith (1982) provided a precise rationale for the motivational role of salient financial incentives in experimental economics.

The second concept of incentivisation is **dominance.** Payoff dominance basically means that the rewards, apart from being salient, have to be large enough

so that the subject cares about which response she is giving. Dominance means that the size of payoffs must be large enough to be *noticeable*. This is called "just-noticeable-difference" in psychology. Smith and Walker (1993) show that subjects respond to increased incentives by expending more cognitive effort, and that this effort can reduce the variance of responses. Of course, reduced variation is different from reduced bias.

One aspect of dominance, which is subtle, is that the payoffs must be sufficiently rewarding to avoid the "flat payoff problem" stressed by Harrison (2006a). The difference in payoffs has to be enough to motivate the subject to engage in some cognitive effort. Imagine, a subject is uncertain about choosing between outcome A and outcome B, however, the payoff difference between changing from choosing A to B is too *unnoticeable*, such as $0.01. Such a small difference is arguably not *dominant* enough for a subject to be motivated to take more cognitive effort to decide on whether to choose one outcome or the other. A flat payoff function means no matter how the allocation of tokens is changed, subjects effectively get the same amount of payoff. A flat payoff function implies that the incentive is not dominant.

The third concept of incentivisation is **incentive compatibility**, assuming salience and dominance. Incentive compatibility means that the incentives are designed in the way that the rewards give the subjects incentives to *truthfully* reveal their preferences or their beliefs. Incentive compatibility of stated choice experiments is often confused with the issue of hypothetical bias. The concept of incentive compatibility is more than just providing real consequences for the choices respondents make. It requires that incentives are designed so that different consequences "make it in the best interest of the respondent to respond truthfully", and "this connection has to be behaviourally transparent and credible" (Harrison 2014, p. 236). In the voting setting, for example, if the voters think that their behaviour will have some impact on the chance that one or the other of two

alternatives will be implemented, and that their utility will be affected by the outcome, they have a positive incentive to behave truthfully and to vote honestly.

To collect the ideas in this section, in my experiment I adopt the QSR to elicit subjects' beliefs. The QSR uses salient and dominant incentives, and it is incentive compatible if the objective is to elicit belief distributions assuming the subjects are risk neutral.

4.4 Summary

In this chapter I provided a critical review of some key experimental studies of social norms conducted by Bicchieri and co-authors, viewed as applications of Bicchieri's philosophical analysis of social norms. I also suggested more sophisticated methodology for experimentally investigating social norms consistent with Bicchieri's account of social norms. In the next chapter, I will introduce the experimental design and tasks applying these methodological protocols, which implement the philosophical account of conditional norms developed by Bicchieri.

Chapter 5 - Experimental Design and Tasks

This chapter will describe the experimental design and the experimental tasks implemented in this thesis. The philosophical analysis of social norms developed by Bicchieri implies that we need two essential experimental tasks: one task in which we allow subjects to make choices that reflect social interaction, and one task in which we elicit subjective belief distributions about behaviours in the first task.

5.1 Experimental Implementation of Bicchieri's Theory of Social Norms

5.1.1 Norm Identification Follow Bicchieri's Theory

Bicchieri (2006, 2017) suggests two steps for norm identification. First, we need to establish that there is *consensus* about what actions are appropriate or inappropriate in specified situations. Bicchieri argues that the mutual consistency of descriptive and normative expectations about behaviour is a necessary condition for identifying a social norm. The second step suggested by Bicchieri is that we must obtain measures to assess the conditions under which a norm will be obeyed. This requires detection of the *conditional preference* for confirming to the norm, given those social expectations.

As argued in chapter 4, the definition of *consensus* deserves closer inspection. To examine Bicchieri's hypothesis on belief consistency, the core step should be rigorous elicitation of both descriptive and normative subjective belief *distributions*. Only then can we test the existence of belief consistency and determine degrees of belief consistency. This requires us to be able to make inferences about the bias and confidence of beliefs.

The experimental design presented here solves two issues identified in previous chapters. The first issue is with the concept of *confidence*. In the context of social norms, the strength of confidence in the beliefs that agents hold are critical to their decisions about norm compliance or norm violation. Bicchieri (2017) argues

that “unless individuals have *confidence* that the reference network, or at least part of it, will support and enact the change, the risk of suffering negative consequences looms large” (p. 94). An example from Bicchieri (2017) is about a wife of a village chief who independently decided to breastfeed her children. This is a behaviour against the social norm in the village at the time. However, because this woman was in a powerful position, her behaviour of breaking the social norm created a wave of imitation by other women in the village. Bicchieri (2017, p. 24) argues that the fact that the woman was in a powerful position increased the *confidence* in the beliefs of other women who imitated her: the confidence that such behaviour would be approved by the relevant reference network (i.e., the other women in the village).

The second issue is that we emphasise the need to elicit *all* four types of beliefs, particularly on the need to elicit 2D and 2N beliefs, consistent with Bicchieri’s philosophical conceptualisation for norm elicitation. These four types of beliefs, following Bicchieri (2017, Table 2.1, p. 70), are *first-order empirical* beliefs about “what others do (empirical expectation)”, *second-order empirical* beliefs about “what others believe I/others do”, *first-order normative* beliefs about “what others should do (personal normative belief)”, and *second-order normative* beliefs about “what others believe I/others should do (normative expectation)”. These four types of beliefs are referred to here as 1D, 2D, 1N, and 2N beliefs, respectively, consistent with the theoretical review in Chapter 2.

The experimental design includes a *strategic-form* Ultimatum Game (UG) and a Belief elicitation task about behaviour in that UG. These two tasks use different subjects from a random sampling from the same population pool. Bicchieri (2006, p. 70) suggests that “we may also ask third parties to observe the results of one or more games, and elicit both their personal beliefs about appropriate behaviour, as well as their second-order beliefs about what most other third parties think is appropriate behaviour (normative expectations)”. Although just referring to what we also call normative beliefs here, this is exactly what is done in the present experimental

design. One should ideally use random samples drawn from the same population, as well as controlling for core demographics variations in finite samples (e.g., by gender).

In this context, the "behavioral rule R in situations of type S", the definition of a social norm according to Bicchieri, refers to the "rejecting behaviour" by Responders in the UG. The two options that Responders are offered are to either Accept or Reject splits of [50%, 50%] or [80%, 20%]. A split of [50%, 50%] means that the Proposer keeps 50% of the endowment for herself and sends the other 50% to the Responder. A split of [80%, 20%] means that the Proposer retains 80% of the endowment for herself and sends 20% of the endowment to the Responder. We ask: "In a situation of type S (UG), where a Proposer receives an endowment, and chooses between a split of [50%, 50%] and [80%, 20%], does a behavioural rule R exist, that is does a social norm exist, that prescribes what the Responders should do?".

The Belief elicitation task implemented separately from the UG elicits the four types of beliefs characterised above.

5.1.2 Why the Ultimatum Game?

There are two reasons for our experiment to employ the UG. First, the UG is one of the most widely employed games in experimental work on social norms (e.g., Bicchieri and Chavez 2010, 2013; Bicchieri and Xiao 2009). Second, the UG is parsimonious in terms of the number of pure strategies that each player has available. This is particularly important, as we will see momentarily, when the test for the presence of a social norm requires belief elicitations for each relevant strategic choice. In the simple UG we use, there will still be 20 belief elicitations, as shown below. Using games with more pure strategies, such as an UG with more than two proposals allowed, or a Trust game, would add needless complexity for present purposes. The simplicity allows us to conduct an empirical application of Bicchieri's

model of norms under the cleanest possible conditions.

5.2 Experimental Design

Subjects recruited in this study were undergraduate students at the University of Cape Town, South Africa. The experiment includes two incentivised tasks, an UG task and a Belief elicitation task. In addition to the incentivised tasks, we administrated two non-incentivised surveys: a demographic questionnaire and a normative values survey question. The experiments are designed to be between-subject, so the two main tasks were separated into two independent sessions with different subjects. The first session, featuring the UG task, was run on November 3, 2021; and the second session, featuring the belief elicitation task, was run on November 23, 2021. In both sessions subjects answered a demographic questionnaire and the normative values survey. All incentives were in South African Rand (ZAR), referred to as "R" in official notation and in the experimental instructions. At the time of the experiment, the exchange rate between the United States Dollar and the ZAR was $1= R15.5567, and the Purchasing Power Parity (PPP) was $1=R7.040.[31] The sessions were conducted online. The experimental software used in our experiment is oTree, developed by Chen, Schonger and Wickens (2016), and our experiment was programmed by Dr. Brian Monroe.

5.2.1 Ultimatum Game

Our experimental design employed a strategic-form UG, referred to in the instructions as a "Proposer-Responder task". It is a variation of the standard UG. In a standard UG there are two roles, referred to as Proposer and Responder. A Proposer is asked to divide an endowment between herself and a (typically anonymous) partner. The recipient ("Responder") can, in turn, accept or reject the

[31] The source for the PPP estimate is https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm.

Proposer's offer. If the Responder accepts the offer, then both the Proposer and the Responder keep the amounts chosen by the Proposer. If the Responder rejects the offer, then both the Proposer and the Responder get nothing.

The *strategic-form* UG applied in our experiment requires all subjects to make decisions as *both* Proposer and as Responder. After they have all made their decisions in both roles, each subject will be randomly and anonymously matched with one of the other participants in the study. After the match is made, a random draw will determine which role each participant is in. If a subject is selected as Proposer, the decision she made in the role as Proposer will be used, and the decision the other participant made in the role of Responder will be applied. The strategic setting of the UG is to have the subjects make decisions first, then get matched by random draw. The random draw determines which role each subject actually takes and determines each subject's earnings.

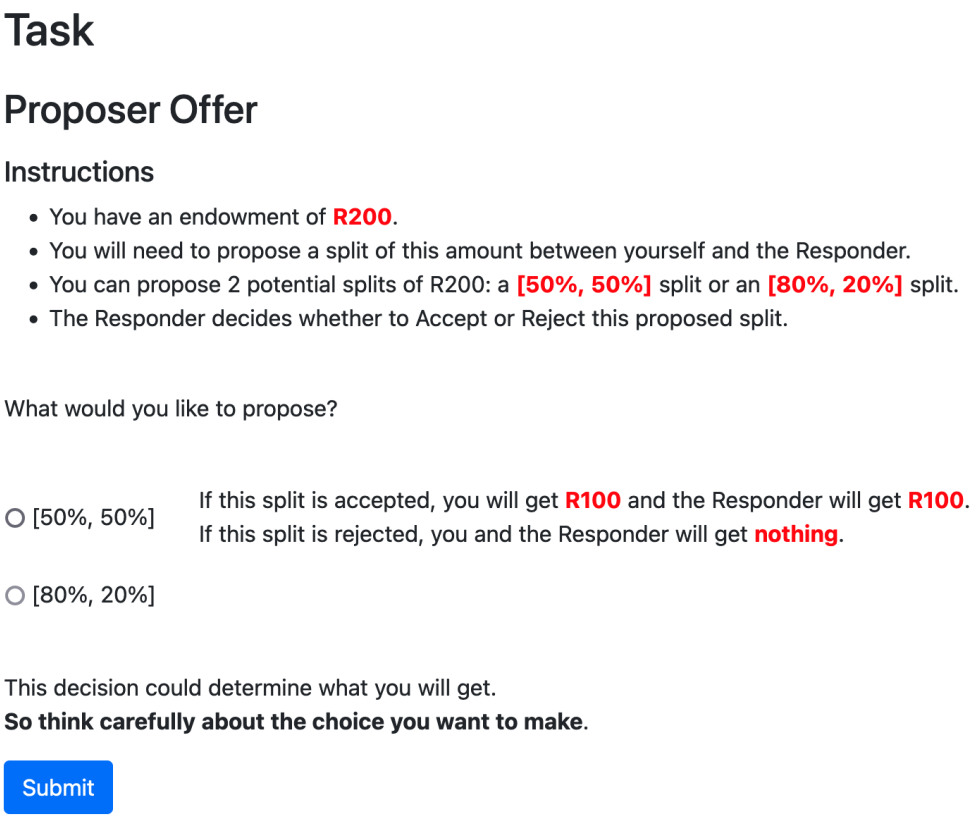
The task instructions applied in our experiment are as follows:

In this task you will be asked to make a decision about dividing an amount of money between yourself and another person. You may also receive money as a result of the decision of another person. In this task there are two roles, referred to as **Proposer** and **Responder**. You and the other participants in this study will be asked to make decisions both as Proposer and as Responder.

After you and all the other participants have made decisions in both roles, you will be randomly and anonymously matched with one of the other participants in the study. All the people participating in this study are UCT students.

After the match is made, a random draw will determine if you will be the Proposer and the other participant the Responder, or the other way around. If you are selected as Proposer, the decision you made in the role of Proposer will be used, while the decisions the other participant made in the role of Responder will be used. The decisions you made in both roles before you were matched, and the random draw that determines which role you will actually take, will determine your earnings.

The UG task works as follows. At the start, in the role of Proposer each subject was given a monetary endowment, either R100 or R300 (randomly determined by the experimental software) as explained to subjects as follows:

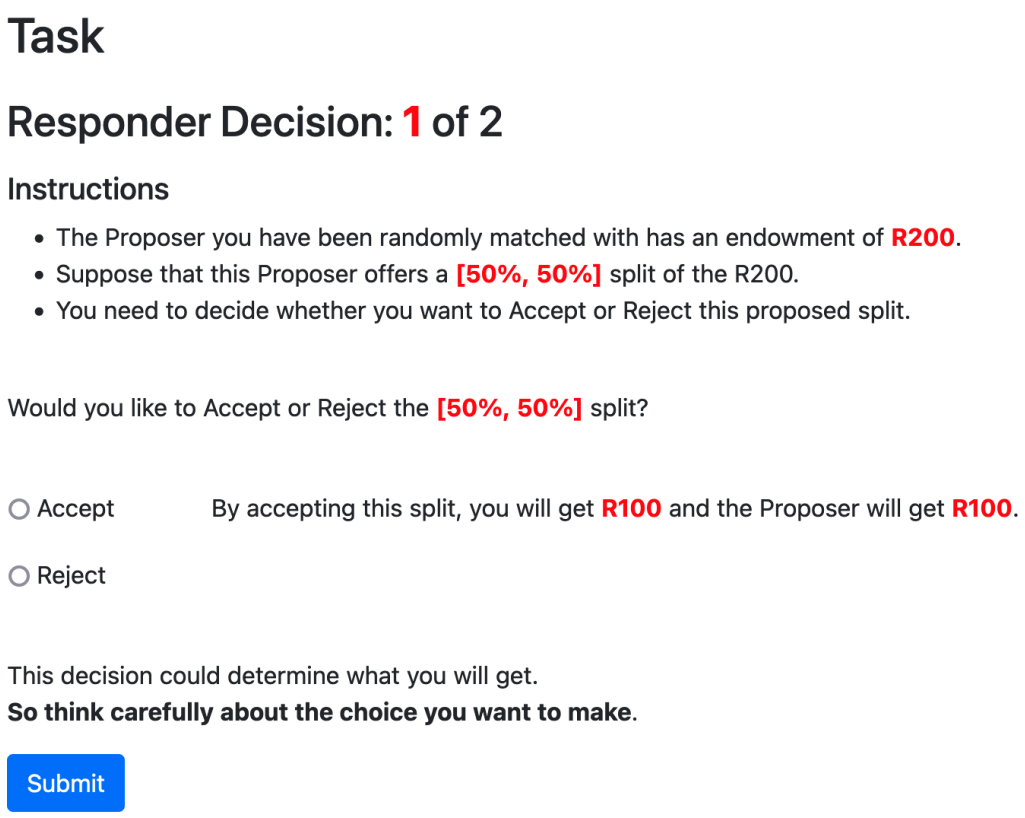


As you can see, the display prompts the Proposer to choose between two offers. The first offer is a [50%, 50%] split of R200, corresponding to the Proposer keeping R100, which is 50% of the endowment, and the Responder receiving R100, which is 50% of the endowment. The second offer is an [80%, 20%] split of R200, corresponding to the Proposer keeping R160, which is 80% of the endowment, and the Responder receiving R40, which is 20% of the endowment. The Proposer must decide what split to offer to the Responder, knowing that the Responder will have made a decision whether to accept or reject the [50%, 50%] offer if it is made, and whether to accept or reject the [80%, 20%] offer if it is made. When the Proposer hovers over a choice of [50%, 50%] or [80%, 20%], the Proposer is shown the amounts that the Proposer and the Responder would earn from this choice, as you can see in the screenshot.

We adopted R200 as the illustrative endowment in the instructions, instead of using the actual endowment amounts of R100 or R300. The software employed in our experiment allowed the subjects to see the amounts that the Proposer and the Responder would earn for each option (i.e., [50%, 50%] and [80%, 20%]) when the Proposer hovers her cursor around each choice. For example, as shown in the screenshot above, when a subject's cursor moved near the option of a split of [50%, 50%], the subject saw a message stating "If this is accepted, you will get R100 and

the Responder will get R100. If this is rejected, you and the Responder will get nothing". If a Proposer's cursor moved near the option of a split of [80%, 20%], she saw a message stating "If this is accepted, you will get R160 and the Responder will get R40. If this is rejected, you and the Responder will get nothing". This helped reduce the cognitive burden for subjects of calculating their gains or losses in the task.

Here are the instructions for the Responder:



As you can see, the display shows that the Responder must decide whether to accept or reject an offer of a [50%, 50%] split. If the Responder accepts the [50%, 50%] split, then the Responder receives R100, which is 50% of the endowment, and the Proposer keeps R100, which is 50% of the endowment. If the Responder rejects the offer, then the Proposer loses the R200 endowment, and both the Proposer and Responder earn nothing. When the Responder hovers over a choice of Accept or Reject, the Responder is shown the amounts that the Proposer and the Responder would earn from this choice, as you can see in the screenshot. The Responder must also decide whether to accept or reject an offer of an [80%, 20%] split. If the Responder accepts the [80%, 20%] split, then the Responder receives R40, which is 20% of the endowment, and the Proposer keeps R160, which is 80% of the endowment. If the Responder rejects the offer, then the Proposer loses the R200

endowment, and both the Proposer and Responder earn nothing.

The next set of instructions illustrate to the subjects how the matching worked and how it determined the earnings of each subject, after they made their choices in the roles of Proposer and Responder.

Once you have made your decisions in both the roles of Proposer and Responder the task is over, and your role as either Proposer or Responder will be randomly determined for payment. At the end of the study, we will determine your earnings for this task in the following way:

- You will be randomly and anonymously matched with another participant in the study. You will not know who this person is and they will not know who you are.
- If you have been assigned the role of Proposer, the person you are matched with will have been assigned the role of Responder. On the other hand, if you have been assigned the role of Responder, the person you are matched with will have been assigned the role of Proposer.
- If you are randomly assigned to the Proposer role, the split you offered in that role will be compared to the Responder's decision to accept or reject that specific split. If the Responder chose to accept that split, each of you will be paid the corresponding amounts. If the Responder chose to reject that split, you lose the R200 endowment and both of you earn nothing.
- If you are randomly assigned to the Responder role, your decisions of whether to accept or reject the offer of a [50%, 50%] split and an [80%, 20%] split will be compared to the split that the Proposer offered. If you, as the Responder, chose to accept this split, each of you will be paid the corresponding amounts. If you chose to reject that split, the Proposer loses the R200 endowment and each of you earn nothing.

We further provided two examples to demonstrate the earnings to the subjects:

For example, suppose that you are randomly selected as Proposer, and you chose to offer the [50%, 50%] split, meaning you keep R100 and the Responder receives R100. Suppose that the person you are anonymously matched with chose to accept this split. Then you earn R100 as Proposer, and the Responder also earns R100. If instead the Responder chose to reject this split, you lose the R200 endowment and both of you earn nothing.

As another example, suppose that you are randomly selected as Responder and that the person you are anonymously matched with in the Proposer role chose to offer the [80%, 20%] split, meaning you receive R40 and the Proposer keeps R160. Assume that as the Responder you chose to accept this split. Then, as Responder, you earn R40, and the Proposer earns R160. If you chose to reject this split, then the Proposer loses the R200 endowment and both of you earn nothing.

Subjects in the session for the UG task were paid a R100 participation fee, in addition to the amounts they earned in the interaction. All subjects answered

demographic questions and completed a normative values survey.

5.2.2 Belief Elicitation Task

In the belief elicitation task subjects were paid according to how accurate their beliefs were about the outcomes of interactions between subjects in the UG session, and how accurate their predictions were of other people's beliefs about these outcomes. The task introduction in the instructions was as follows:

> This is a task where you will be paid according to how accurate your beliefs are about the outcomes of an interaction between people, and how accurate your predictions are of other people's beliefs about these outcomes. Your earnings will depend on what the outcomes of the interaction between other people actually are, and on what other people report that they believe about the interaction. You will be presented with some questions and asked to place bets on your beliefs about the answer to each question. You will be rewarded for your answer to one of these questions, so you should think carefully about your answer to each question. The question that is chosen for payment will be determined after you have made all decisions, and that process is described at the end of these instructions. Everyone participating in this study is a UCT student.

The information above intends to inform the subjects in our belief elicitation task what the reference network was for the subjects who participated the UG task prior to them.

In the belief elicitation session, subjects were first shown how the UG task (Proposer-Responder task) in the previous session worked, as per the instructions below:

> The interaction between people, on which you will be asked to place bets, works as follows:
>
> - There are two roles in the interaction: **Proposer** and **Responder**
> - The Proposer is given a money endowment of R100, and is asked to propose a split of this amount between themself and the other person, the Responder.
> - The split the Proposer can offer is either [50%, 50%] or [80%, 20%] of the R100 endowment, where the first percentage in each potential split is the percentage of the endowment the Proposer would get, and the second percentage in each potential split is the percentage of the endowment the Responder would get. Therefore, with a [50%, 50%] split, the Proposer gets R50 and the Responder gets R50. With an [80%, 20%] split the Proposer gets R80 and the Responder gets R20.

- The Responder will then be asked to decide, for each potential split of the money that the Proposer might offer, whether to accept or reject this proposed split.
- If the Responder **accepts** a proposed split, the Responder receives the amount they were offered, and the Proposer keeps the rest of the money.
- If the Responder **rejects** a proposed split, the R100 endowment is withdrawn, and both the Proposer and the Responder get nothing.

Thus, the Proposer has to make one choice: to propose either a [50%, 50%] split of R100, or an [80%, 20%] split of R100.

By contrast, the Responder has to make two choices: to accept or reject each potential split. If the Responder accepts a proposed [50%, 50%] split of R100, the Proposer gets R50 and the Responder gets R50. If the Responder rejects this split, both the Proposer and the Responder earn nothing from that interaction. If the Responder accepts a proposed [80%, 20%] split of R100, the Proposer gets R80 and the Responder gets R20. If the Responder rejects this split, both the Proposer and the Responder earn nothing from that interaction.

In the instructions we also informed the subjects in the session for the belief elicitation task that 255 UCT students took part in the UG task. These subjects in the belief elicitation task were also shown screenshots which were seen by the subjects who participated in previous session for the UG task, in order to ensure that the subjects in the belief elicitation session understood fully how the UG task worked.

Next, we showed the subjects in the current task how the belief elicitation task worked, with instructions summarising the four different types of belief questions following the experimental design:

In this task you will be asked for your beliefs about the behaviour of others when they are in the *Responder* role, and also for your predictions of what other people in this study believe about these outcomes. There are four types of questions: —**Type 1 -** What do you believe UCT students **actually did** in the Responder role in response to the potential splits of [50%, 50%] and [80%, 20%]? —**Type 2 -** What do you predict the **other people** completing this task today believe about what UCT students **actually did** in the Responder role? —**Type 3 -** What do you believe UCT students in the Responder role **should have done** in response to the potential splits of [50%, 50%] and [80%, 20%]? —**Type 4 -** What do you predict the **other people** completing this task today believe about what UCT students **should have done** in the Responder role?

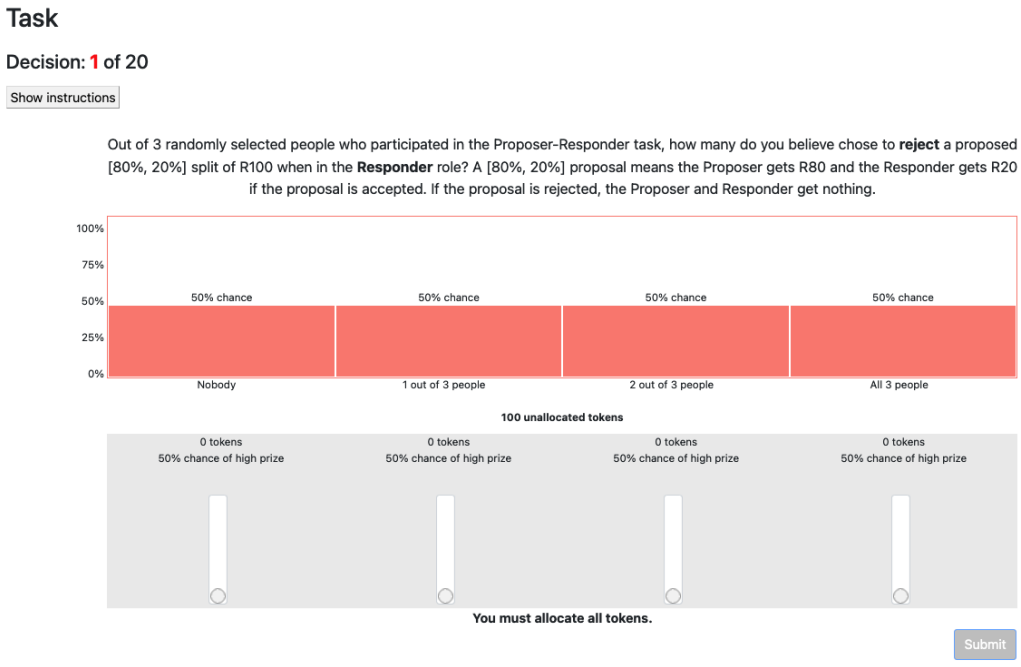
The instructions then illustrated each type of question with the following screenshots describing of the question types:

Question Type 1

The first type of question is about your beliefs concerning what UCT students **actually did** in the Responder role in response to the potential splits of [50%, 50%] and [80%, 20%]. For example, you will be asked "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [50%, 50%] split of R100 when in the **Responder** role?". You will also be asked "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". You will need to allocate 100 tokens to express your beliefs about the possible answers to each of these two questions.

Our Question Type 1 and Type 3 ask about *3 people*. This is because we prefer to design the social context to be more than just one other person, but not so many that there were too many second-order belief questions (2D and 2N) for Question Type 2 and Type 4.

The screenshot below shows Question Type 1 for the case of an [80%, 20%] split:



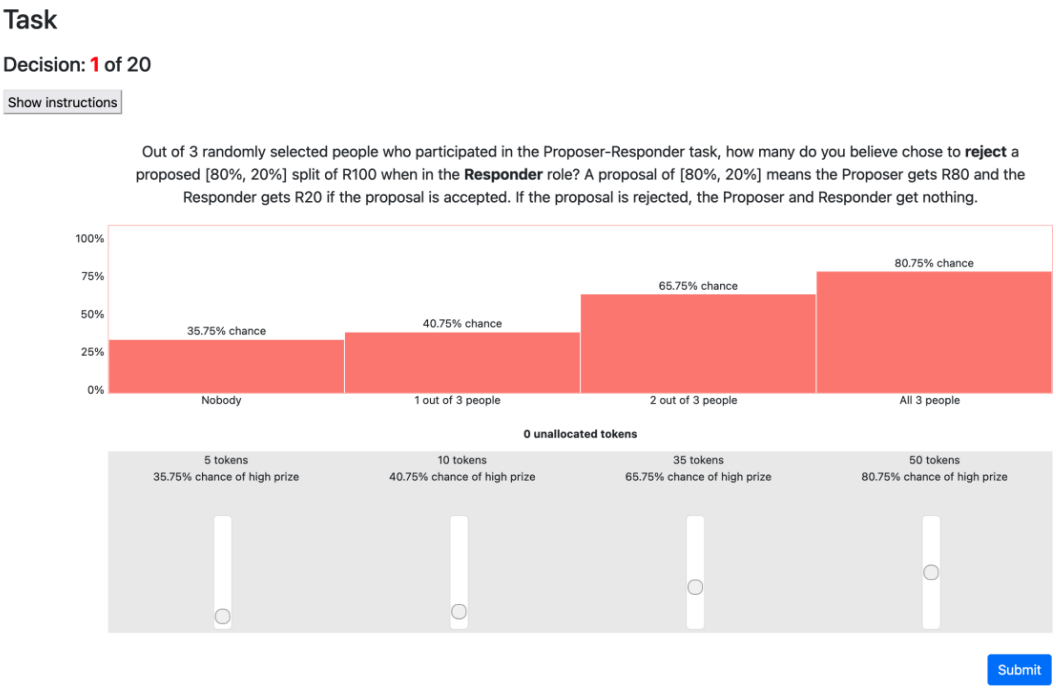
Subjects had 4 sliders to adjust in this task, shown at the bottom of the screen, and 100 tokens to allocate across the sliders. Each slider allowed subjects to allocate tokens to reflect their beliefs about the answer to this question. Subjects were required to allocate all 100 tokens, and the software always started with 0 tokens

allocated to each slider. As subjects allocated tokens, by adjusting sliders, the percentages displayed on the screen do change. The potential earnings of a subject were based on the percentages that were displayed after they had allocated all 100 tokens, where higher percentages mean a higher chance of receiving a larger prize of R500 as opposed to a smaller prize of R50. The instructions demonstrate how the interface worked:

Where you position each slider depends on your beliefs about the correct answer to the question. The bars above each slider correspond to that particular slider. In our example, the tokens you allocate to each bar will naturally reflect your beliefs about the question, "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". The first bar corresponds to your belief that **Nobody** chose to reject a proposed [80%, 20%] split of R100. The second bar corresponds to your belief that **1 out of 3 people** chose to reject a proposed [80%, 20%] split of R100. The third bar corresponds to your belief that **2 out of 3 people** chose to reject a proposed [80%, 20%] split of R100. Finally, the fourth bar corresponds to your belief that **All 3 people** chose to reject a proposed [80%, 20%] split of R100. Each bar shows your percentage chance of earning R500 as opposed to R50, depending on what 3 randomly selected UCT students **actually did** in the Proposer-Responder task.

The instructions provided an example to illustrate how subjects were to use the sliders to reflect their beliefs by imagining some actual numbers:

Suppose you are answering the question we just discussed, "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". Suppose you think that out of 3 randomly selected UCT students who took part in the Proposer-Responder task, there is a good chance that "All 3 people" in the Responder role chose to reject a proposed [80%, 20%] split of R100. Then you might allocate 50 tokens with the slider for "All 3 people". Suppose you also think there is a pretty good chance that "2 out of 3 people" in the Responder role chose to reject a proposed [80%, 20%] split of R100. Then you might allocate 35 tokens with the slider for "2 out of 3 people". Finally, suppose you think there is a low chance that "1 out of 3 people" in the Responder role chose to reject this proposal, and an even lower chance that "Nobody" in the Responder role chose to reject this proposal. Then you might allocate 10 tokens with the slider for "1 out of 3 people," and 5 tokens with the slider for "Nobody". This is what the display would look like if those were your choices:



In the example, because all 100 tokens have been allocated, the **Submit** button becomes clickable so that you can submit your choice and move on to the next question. If you would like to change your token allocation before clicking the Submit button, then just use the sliders to make any adjustments. For example, you could allocate zero tokens with one or more of the sliders.

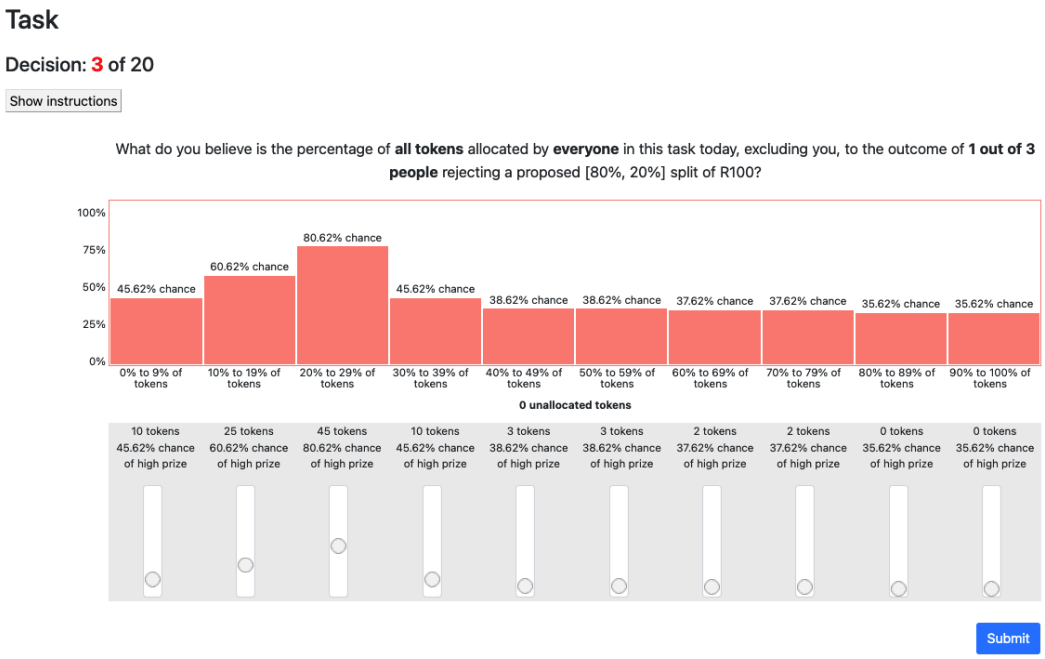
Question Type 2 is about what subjects predicted the "other people completing this task today" believed about what UCT students *actually did* in the Responder role. In other words, subjects were predicting how the other people completing this task actually allocated their 100 tokens for Question Type 1. Again, subjects were asked to allocate 100 tokens to express their beliefs about the possible answers to these questions, but in this case subjects had 10 sliders to adjust to reflect their beliefs about the answer to each question. This is what the set of questions were:

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **Nobody** rejecting a proposed [80%, 20%] split of R100?
- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **1 out of 3 people** rejecting a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **2 out of 3 people** rejecting a proposed [80%, 20%] split of R100?
- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **All 3 people** rejecting a proposed [80%, 20%] split of R100?

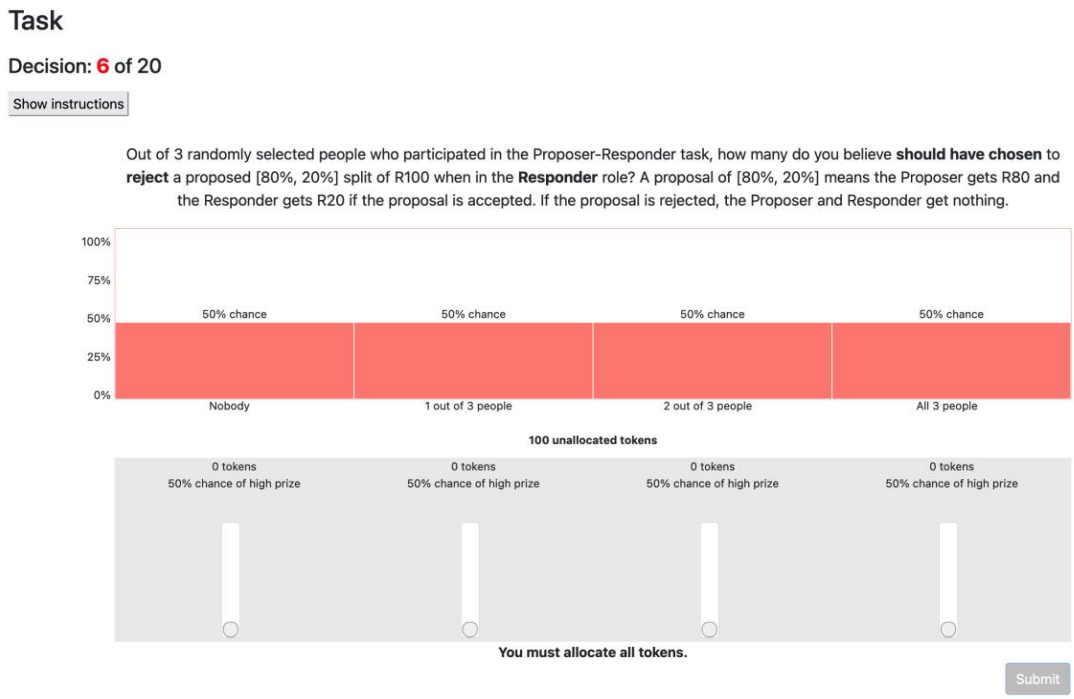
The instructions illustrate Question Type 2 by the following example:

Once again, we'll imagine some actual numbers that you might not think are very likely. But, as before, they are just for the sake of this example. Suppose you are answering the question, "What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **1 out of 3 people** rejecting a proposed [80%, 20%] split of R100?". Suppose that you predict there is a good chance that the people completing this task today allocated 20% to 29% of all tokens to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate 45 of your tokens with the slider for "20% to 29%". Perhaps you predict there is also a fairly good chance that the people completing this task today allocated 10% to 19% of all tokens to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate 25 tokens with the slider for "10% to 19%". Perhaps you think there is an equal chance that people allocated either 0% to 9% or 30% to 39% to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate 10 tokens with the slider for "0% to 9%", and 10 tokens to the slider for "30% to 39%". Finally, suppose you think there is a very low chance that more than 40% of all tokens were allocated to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate your remaining 10 tokens as shown in this screenshot.



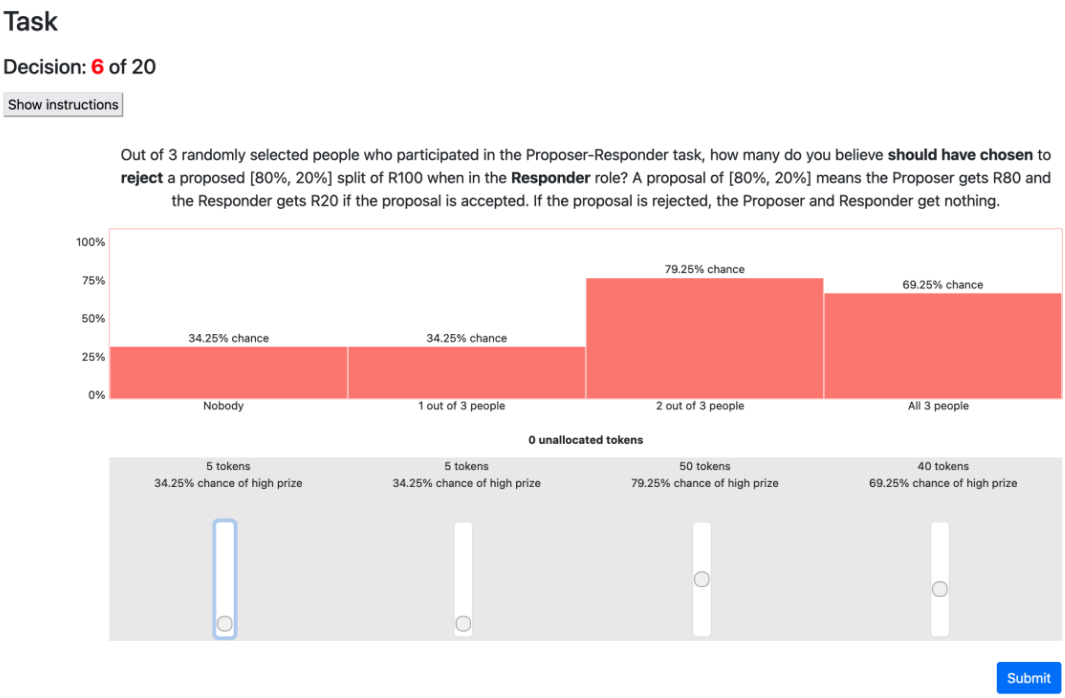
Remember you are free to allocate zero tokens with one or more of the sliders, just as we did for sliders 80% to 89%, and 90% to 100% in the screenshot.

Question Type 3 is about what subjects in this session believed UCT students in the Responder role in the previous UG task session *should have done* in response to the potential splits of [50%, 50%] and [80%, 20%]. For example, subjects were asked, “Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe *should have chosen* to *reject* a proposed [80%, 20%] split of R100 when in the **Responder** role?”. Subjects needed to allocate 100 tokens to express their beliefs about these two questions: one question for the potential [50%, 50%] split, and one question for the potential [80%, 20%] split. These questions were not incentivised. However, the Type 4 questions, which *were* used to determine their payment, were about how other people allocated their tokens to the Type 3 questions. This screenshot shows what Question Type 3 looks like.



The instructions provide an example for the demonstration of Question Type 3:

Let's look at another example. Once again we'll imagine some actual numbers that you might not think are very likely. But, as before, they are just for the sake of this example. Suppose you are answering the question, "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe **should have chosen** to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". Suppose you think that out of 3 randomly selected UCT students who participated in the Proposer-Responder task, **at least** 2 out of 3 people **should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 50 tokens with the slider for "2 out of 3 people", and 40 tokens with the slider for "All 3 people". Suppose you think that the sliders for "Nobody" and "1 out of 3 people" are equally likely. Then you would allocate 5 tokens with the slider for "Nobody" and 5 tokens with the slider for "1 out of 3 people". This is what the display would look like if those were your choices:



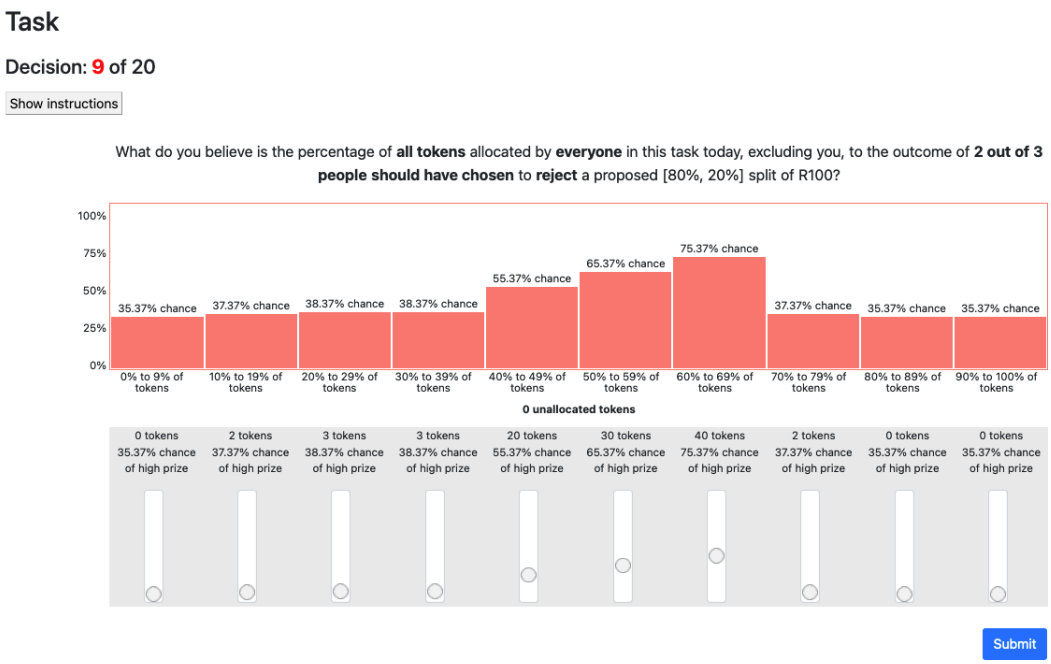
Question Type 4 is about what people in this session predict the other people completing this task today believe about what UCT students *should have done* in the Responder role. In other words, subjects were asked to predict how the other people completing this task today actually allocated their 100 tokens for Question Type 3. Again, each subject was asked to allocate 100 tokens to express their beliefs about the possible answers to these questions, but in this case each of them had 10 sliders to adjust to reflect their beliefs about the answer to each question. In Question Type 4 subjects were *incentivised* to predict responses to the *non-incentivised* Question Type 3. This is what the set of questions look like:

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **Nobody should have chosen** to **reject** a proposed [80%, 20%] split of R100?
- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **1 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?
- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **All 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?

Next is one example out of the set of 4 questions for Question Type 4 applied in the instructions:

Let's look at another example. Once again, we'll imagine some actual numbers that you might not think are very likely. But, as before, they are just for the sake of this example. Suppose you are answering the question, "What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?". Suppose that you predict there is a good chance that the people completing this task today allocated 60% to 69% of all tokens to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 40 of your tokens with the slider for "60% to 69%". Perhaps you predict there is also a pretty good chance that the people completing this task today allocated 50% to 59% of all tokens to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 30 tokens with the slider for "50% to 59%". Perhaps you think there is also a fairly good chance that people allocated 40% to 49% of all tokens to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 20 tokens with the slider for "40% to 49%". Finally, suppose you think there is a very low chance that less than 40% or more than 70% of all tokens were allocated to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate your remaining 10 tokens as shown in this screenshot.



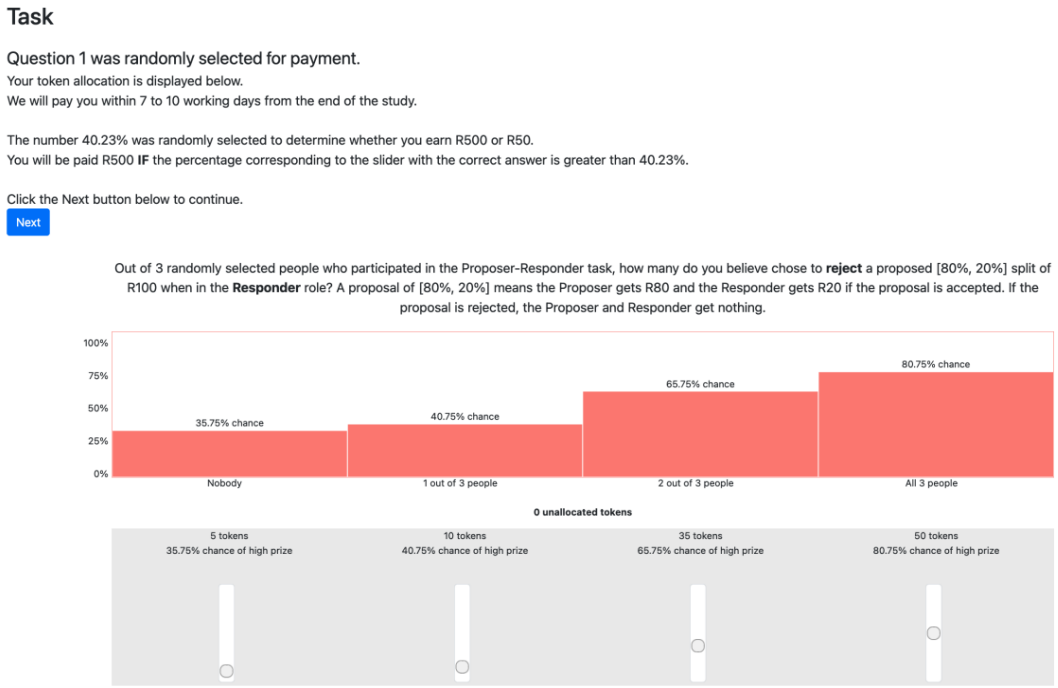
Subjects in this session were paid for one randomly chosen question among their answers to the questions of Type 1, 2 or 4. Following our experimental design of applying the binary lottery payment procedure so as to induce risk neutral responses, the prizes in this task were either R500 or R50. The instructions explain how responses to the randomly selected question determined earnings for the subject in this task:

As mentioned earlier, by allocating tokens with the sliders the percentages on the screen change. These percentages are between 0 and 100, and they represent the probability of winning different money prizes depending on the allocation of your tokens. The prizes are either R500 or R50. The higher the percentage for a slider, the greater your chance of being paid R500 instead of R50. On the other hand, the lower the percentage for a slider, the smaller your chance of being paid R500 instead of R50. If you allocate all your tokens to one slider this gives you a 100% chance of being paid R500, if the correct answer is represented by that slider.

To determine payment for this task, the computer will randomly select one question of type 1, 2 or 4. The decision screen selected will be shown back to you and the computer will record the percentages you received from allocating your tokens. You will either be paid R500 or R50 depending on your token allocation and the correct answer to the randomly selected question.

For further explanation of the payment procedure, the instructions included an example supposing that a question of Type 1 had been randomly selected for

payment:



Suppose the question (for your payment) that gets randomly selected is, “Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?”. Finally, suppose you allocated 35 tokens with the slider for “2 out of 3 people” and out of the 3 randomly selected UCT students, 2 of them actually did reject this proposal. With 35 tokens allocated to this answer you have a 65.75% chance of winning R500, and therefore a chance of 34.25% of winning R50, where 34.25% is just 100% - 65.75%. The computer will then randomly draw a number between 0 and 100, with every number between 0 and 100 being equally likely. If the randomly drawn number is less than or equal to the percentage you received for the correct answer, you will win R500. If the randomly drawn number is greater than the percentage you received, you will win R50.

Thus, your earnings depend on your reported beliefs and, of course, the true answer. For example, if the percentage you received is 65.75%, as in our example, and the number that is randomly drawn is 40.23%, you will be paid R500 since 40.23% is less than 65.75%. This randomly drawn number of 40.23% is shown in the screenshot. However, if the randomly drawn number is 79.71%, because this is greater than 65.75% you will be paid R50. Finally, if you allocate all of your tokens to one slider that corresponds to the correct answer this means you will be paid R500 with certainty, because a randomly drawn number between 0 and 100 will always be less than or equal to 100.

5.2.3 Demographics Questionnaire and Normative Values Survey

In order to control homogeneity to some degree across our subject sample, we adopted a demographic questionnaire and designed a normative values survey. We can use responses to these surveys to better control for a Normative Reference Network across our subjects.

The demographic questionnaire applied in our experiment is the standard one used in the RUBEN/CEAR-Africa Lab, and the normative values survey is a selection of questionnaires from Afrobarometer questionnaires.[32] In the demographic questionnaire, we asked each subject to report their age, gender, geographical location, marriage status, income, financial status, race, and their willingness to take risks (scale from 0 to 10).

One way to judge if the sample constitutes a Normative Reference Network is to judge if the sample reports "consensus" about support for a series of value statements. We therefore designed the normative values survey, which asked each subject to report to what extent they agree with or disagree with a series of normative statements included in the survey. Each subject was asked to respond between 1 to 10 to indicate agreement or disagreement with each statement. The specific questions posed to each of our subjects were as follows, with our acronym listed first in bold:

1. **ReligLaw**: Religious beliefs towards the laws (1 is support, 10 is against)
2. **ServTax**: Improving government services with taxes (1 is support, 10 is against)
3. **EqualLnd**: More equal distribution of wealth (1 is support, 10 is against)
4. **OrderLib**: Maintaining law & order or defending civil liberties (1 is law & order, 10 is defending civil liberties)
5. **PrivPub**: Privatising public enterprises (1 is support, 10 is against)

[32] Afrobarometer questionnaires are applied widely across Africa to explore the breadth of African public opinions. More information can be found at https://www.afrobarometer.org/surveys-and-methods/

6. **Abortion**: Keeping abortion illegal (1 is support, 10 is against)
7. **Compete**: Competing with fellow workers (1 is support, 10 is against)
8. **Particip**: Citizen participation in government decision-making (1 is support, 10 is against)
9. **Community**: Putting community well-being ahead of own interests (1 is support, 10 is against)
10. **OurWay**: Defending our way of life instead of becoming more like other countries (1 is defending, 10 is becoming more like others)
11. **GovParent**: Government treating citizens like their children rather than as the boss (1 is like children, 10 is like the boss)
12. **Conflict**: Avoiding conflict at all costs (1 is support, 10 is against)
13. **TimeResolves**: Resolving problems with time, rather than as soon as we can (1 is support, 10 is against)
14. **GovWell_Being:** Government should bear the main responsibility for well-being of people (1 is support, 10 is against)
15. **Customs**: Better for society if ethnic groups maintain their customs & traditions (1 is support, 10 is against)
16. **COVID**: More resources devoted to COVID, even if less for other government services (1 is support, 9 is against)

We do not need to have subjects hold particular views on these issues.

Our measure of consensus is derived and explained by Tastle and Wierman (2007), and is intended to reflect several criteria for consensus using ordered Likert scales such as the ones we used. Some measures, such as Entropy Measures, do not take into account that equivalent positive sample responses of 1 and 2 for a particular survey question reflect more consensus than the same sample responses of 1 and 10. To account for this notion of similarity one must model the values of the Likert responses. There are several ways to do this, and the simplest is just to use the numbers provided to the subjects. As one might hope, arbitrary positive affine

transformations of these numbers have no effect on the consensus score. All of these measures reflect the sample as a whole, and provide no insight into whether any particular subject shared the consensus with others.

5.2.4 Treatments

The experiment included four treatments: an endowment effect treatment, a landing page treatment, a cartoon treatment, and an order effect treatment.

Endowment Effect Treatment

We evaluated two endowment levels for the Proposer in the UG task session. Each subject assigned in the role of Proposer was given a monetary endowment of R100 or R300. The amount each subject received was randomly determined by the experimental software.

This treatment was designed to check if the incentives to engage in certain patterns of rejection behaviour in the UG depend on the opportunity cost of the behaviour. It could be argued that very low stakes would fail to elicit a social norm because the incentives for norm-consistent behaviour, particularly where rejection of [80%, 20%] offers is concerned, are not great enough in terms of monetary consequences compared to alternative norm-inconsistent behaviour. Varying the endowment allows us to evaluate this potential confound for our test of the existence of norms.

An implication of this treatment is also that it allows an evaluation of the amount of income that subjects are willing to forego to enforce norms. This implication allows some evaluation of the willingness to pay to enforce any norm that arises, if it arises. This willingness to enforce a norm is a key feature of Bicchieri's concept of a social norm.

Landing Page Treatment

The landing page treatment was applied in the session for the belief elicitation task. It served the purpose of reminding subjects that Question Type 2 had asked them to predict the answers to questions of Type 1, which they had answered themselves, and that Question Type 4 had asked them to predict the answers to questions of Type 3, which they had also answered themselves, and also to remind the subject of their own responses. When subjects were asked to answer the set of 4 questions of Type 2, we provided information about their own responses before each question of Type 2. Similarly, subjects were also shown information about their own responses before each question of Type 4. This is the form of the information provided, where W, X, Y and Z were filled in by the software:

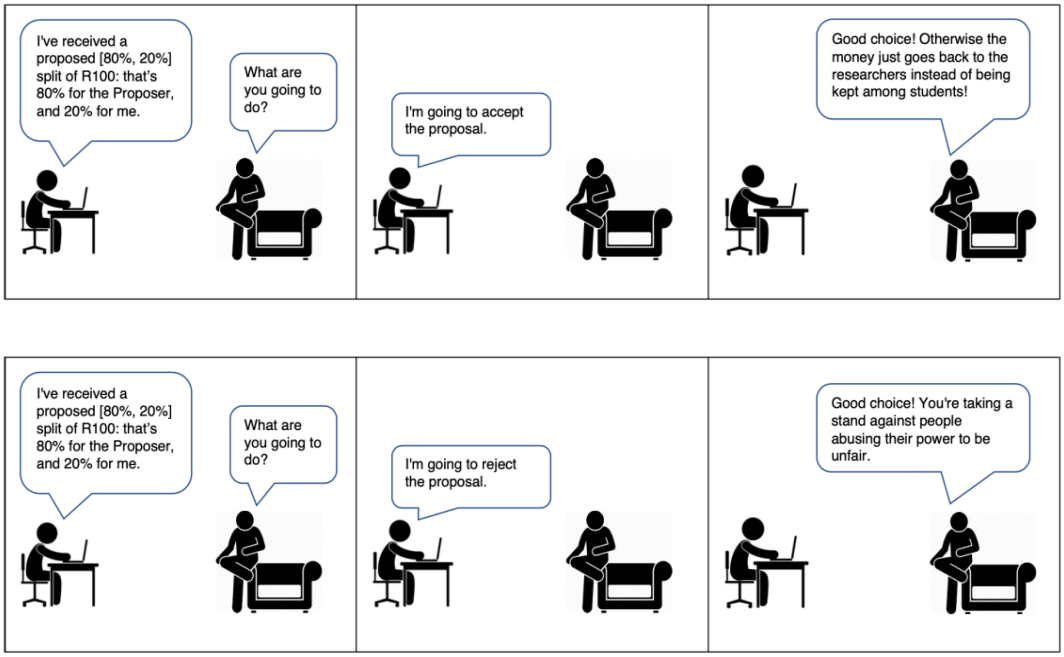
[You allocated W% of your tokens to **Nobody**, X% of your tokens to **1 out of 3 people**, Y% of your tokens to **2 out of 3 people**, and Z% of your tokens to **All 3 people**.]

In our experiment, a randomly selected half of the subjects were provided with this information.

The purpose of this treatment was to acquire information on the extent to which subjects displayed consistency between their 1D answers and their 2D answers simply because they inferred beliefs about others' behaviour from their own case, as opposed to making independent personal and social estimations. If subjects were doing that, then the treatment should allow us to observe a statistical difference between subjects who were reminded about their own responses and subjects who were not reminded.

Cartoon Treatment

For Question Type 3, we suspect that subjects might find it strange that someone could believe that different people *should* make different choices in this situation. These cartoons are designed to help our subjects to see how there could be variation in what people believe other people *should* choose:



There are precedents in the experimental economics and survey literature (e.g., Harrison 2006b) for showing subjects rationales for alternative answers to questions. Of course, with all such treatments care must be taken not to steer subjects toward one alternative. Emphasis is thus on "showing" rather than "telling", and cartoons have sometimes been used for this purpose (Harcourt-Cooke, Els and Van Rensburg 2022).

The cartoon treatment is motivated by two concerns we had when designing the experiment.

The first concern is that some subjects might be misled by semantic ambiguity in everyday uses of the word 'should'. English speakers often use locutions such as "There should be lots of people coming to the rally just to be out in the sun". This is a *predictive* rather than a *normative* use of 'should'. In the case of our belief elicitation task, this ambiguity could cause some subjects to think that 1D and 1N questions, and 2D and 2N questions, respectively, are querying the same thing. The cartoon is intended to signal to subjects that 'should', in N-type questions, is intended *normatively*.

The force of this first worry is compounded by our second concern. Some subjects might have difficulty seeing how a normative judgment in a setting such as

the UG game could apply less than universally, especially when, as in our experiment, all subjects are in symmetrical situations. This is particularly likely if they think of relevant norms here as having moral force, and they have folk intuitions of a loosely Kantian kind (Kant 1785), as many people do. Such subjects might then *either* be led to interpret 'should' predictively, so as to make better sense of the form of the question, *or* think that "interior" response options (reporting that some subjects should do different things from one another, or that some subjects would believe that other subjects should do different things from one another) are 'tricks', and avoid these responses. The cartoon reminds subjects, without steering them toward any particular answer through direct instruction, that more than one potential norm might guide UG responders. Specifically, they are reminded that some subjects might think that the choice is regulated by a norm of fairness between UG players, while others might think that the relevant norm is for players to act as a team and avoid "wasting" money by letting the experimenters reclaim it.

Order Effect Treatment

There are 20 belief questions in total in the session for the belief elicitation task. They are arranged as follows:

- 2 questions on the first-order *descriptive* beliefs for the 2 proposals ([50%, 50%] split and [80%, 20%] split);
- 2 sets of 4 questions on the second-order *descriptive* beliefs (about the first-order *descriptive* beliefs);
- 2 questions on the first-order *normative* beliefs for the 2 proposals ([50%, 50%] split and [80%, 20%] split);
- 2 sets of 4 questions on the second-order *normative* beliefs (about the first-order *normative* beliefs);

The number of belief questions is therefore $2+2\times4+2+2\times4 = 20$. We designed an order effect treatment in which half of the subjects answered 1D and 2D questions

before they answered 1N and 2N questions; and the other half answered the questions in the opposite order. The experimental software randomly assigned each subject to one of the orders.

This treatment provides a simple check for whether the second set of beliefs of subjects are just replicas of the first set of beliefs. This check is only valid for between-subject comparisons with this treatment. If the subjects responding to the D (1D and 2D) questions *first* generate the "same" responses as those subjects responding to the D questions *second*, then we have some confidence that the *latter* are *not* driven by the responses to the first set of N (1N and 2N) questions. Similarly for comparisons of the first set of N questions and the second set of N questions.

The reason this treatment matters is because we do not want the null hypothesis of belief consistency to be satisfied by the N questions always coming second but being heuristically anchored to the first responses. We want the test of the null hypothesis to be based on the choices of the subjects in both the D and N settings, free of the effect of such possible anchoring.

5.3 Summary

This chapter presented the experimental design and experimental tasks implemented in the thesis. The experiment utilises the methodological protocol introduced in chapter 4. We designed four tasks, two incentivised tasks and two unincentivised tasks, and ran two sessions. The first session employed an incentivised strategic-form UG, and two non-incentivised surveys. The second session conducted an incentivised belief elicitation task, with two non-incentivised surveys. The experiment also included a few treatments.

Chapter 6 - Results

This chapter presents results from the experiment described in chapter 5, and the hypothesis tests listed in chapter 2.

6.1 Description of Observed Behaviour

This section provides a description of the observed behaviour from the two experimental sessions in our study: the session with the UG task and the session with the belief elicitation task.

6.1.1 Ultimatum Game Data

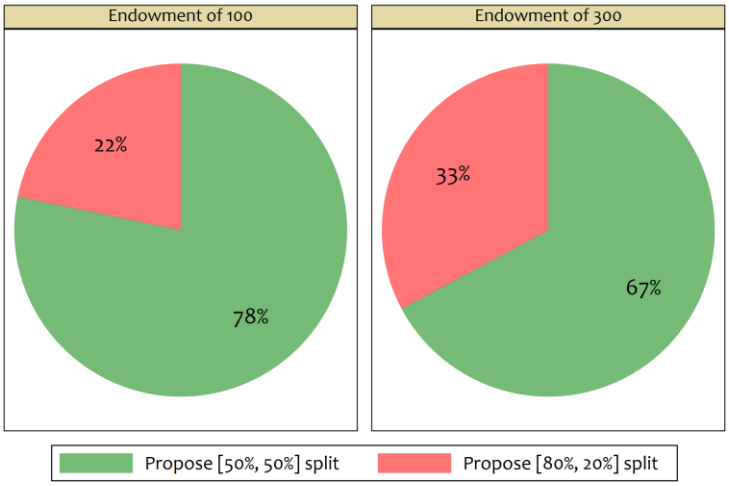
The UG task in our experiment was run online on November 3, 2021, and 255 subjects from the University of Cape Town participated. The total payments for this task were R49,260, for an average of R193 across all subjects. The average official exchange rate in November 2021 was USD $1 = R15.5567, so the total payments were $3,167 with average payments were $12.42. The PPP of the Rand has been estimated to be $1=$7.0400 in 2021; in terms of PPP, the total payments were $6,997 with average payments of $27.44.

Figures 6.1, 6.2 and 6.3 show the UG behaviour that is the object of the beliefs being elicited. Our endowment effect treatment was implemented by randomisation: 124 subjects received an endowment of R100, and 131 subjects received an endowment of R300.

Figure 6.1 shows a comparison of the fraction of the two proposals, the [50%, 50%] split versus the [80%, 20%] split, varying between the two endowments. Figure 6.1 shows that with a higher endowment at R300, less subjects proposed the split of [50%, 50%] and hence more subjects proposed a split of [80%, 20%]. This alerts us to a possibility of an endowment effect. The endowment effect will be formally evaluated statistically later, since Figure 6.1 has no controls of other possible factors which may be correlated to the UG behaviour described here, such

as gender differences across randomised samples.

Figure 6.1: Fraction of Proposals

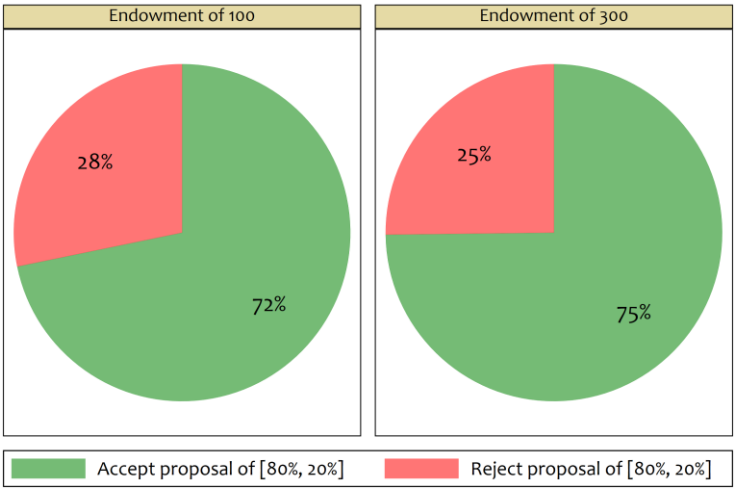


Figures 6.2 and 6.3 show the fraction of responses of the behaviour from Responders' role in the UG task, again varying between the two endowments.

Figure 6.2 shows that the rejection rate for the proposal of [50%, 50%] across the two endowments are the same. However, the rejection rates for the proposal of [80%, 20%] are different across the two endowments, as shown in Figure 6.3. Figure 6.3 demonstrates that when the endowment is higher, there was slightly less rejection behaviour in the Responder role towards the split of [80%, 20%].

Figure 6.2: Fraction of Responses to [50%, 50%] Proposal

Figure 6.3: Fraction of Responses to [80%, 20%] Proposal



6.1.2 Beliefs Data

The belief elicitation task in our experiment was run on November 23, 2021. This task is completely distinct from the UG task. There were 396 subjects in this task, and 390 of them were eligible for payment (the final 6 subjects did not finish the task on time, and are therefore not eligible for payment). The total payments for this task were R132,300, for an average of R340, so total payments were $8,505 with average payments of $21.81. In terms of PPP total payments were $18,793, with average payments of $73.70.

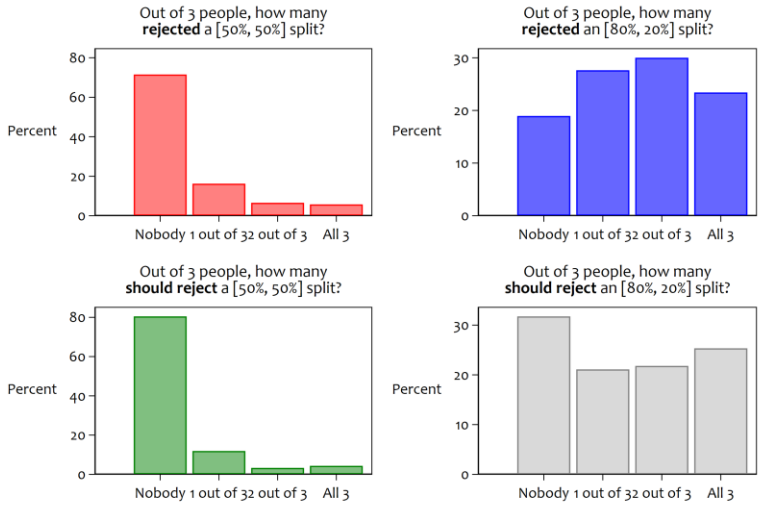
The experimental design for this task includes 4 treatments: an endowment effect, an order effect, a cartoon effect, and a landing page effect, which are randomly assigned to individual subjects by the experimental software. The belief questions about behaviour with an endowment of R100 was used for 172 subjects, and the belief questions about behaviour with an endowment of R300 was used for the remaining 224 subjects. In terms of the order effect treatment, 204 subjects were asked about their beliefs about the [50%, 50%] split first, and 192 subjects were asked about the [80%, 20%] split first. The cartoon treatment was randomly assigned to 294 subjects. The landing page treatment to encourage consistency was provided to 194 subjects.

6.1.2.1 *First-Order Beliefs Data*

The two panels on the left side of Figure 6.4 show a comparison between the 1D and 1N beliefs for the [50%, 50%] split. The two distributions show the same modal belief at bin #1, which refers to the first outcome "Nobody". However, the two distributions vary at each outcome.

The differences in beliefs for each outcome when comparing the 1D and 1N beliefs are more visually apparent for the [80%, 20%] split, as seen in the two panels on the right side of Figure 6.4. First, the modal beliefs of the two distributions are different. In the 1D belief distribution, the mode is at bin #3, which refers to the outcome "2 out of 3 people"; in the 1N belief the mode is at bin #1, which refers to the outcome of "Nobody". In addition, the token allocations again vary across the two distributions for each outcome.

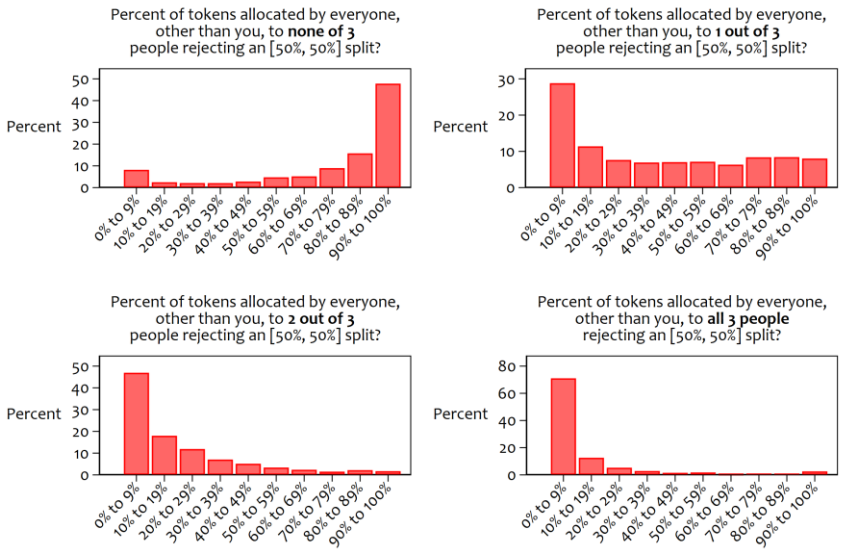
Figure 6.4: First-Order Beliefs



6.1.2.2 *Second-Order Beliefs Data*

Figure 6.5 shows four panels of the 10-bin distribution, which are the 2D second-order descriptive beliefs for the [50%, 50%] split. There are four panels, because each panel corresponds to one of the four outcomes in the first-order descriptive belief question for the [50%, 50%] split displayed in Figure 6.4.

Figure 6.5: Descriptive Second-Order Beliefs for [50%, 50%] Split



For reference, the first-order belief which corresponds to the 4 panels in Figure 6.5 is the top left panel of Figure 6.4, shown here as Figure 6.6. Recall in our experimental design that the relation between the 1D question and the 2D questions is that the 2D questions are about what subjects predicted how the other people in the same experimental task answered 1D question about the same proposal (in this instance the [50%, 50%] split), and how they actually allocated their 100 tokens for each outcome in the 1D question. Hence Figure 6.5 shows beliefs about the behaviour in Figure 6.6, which is just an extract from one of the panels in Figure 6.4, the panel in Figure 6.4 using red for displaying the data.

Figure 6.6:
Out of 3 people, how many **rejected** a [50%, 50%] split?

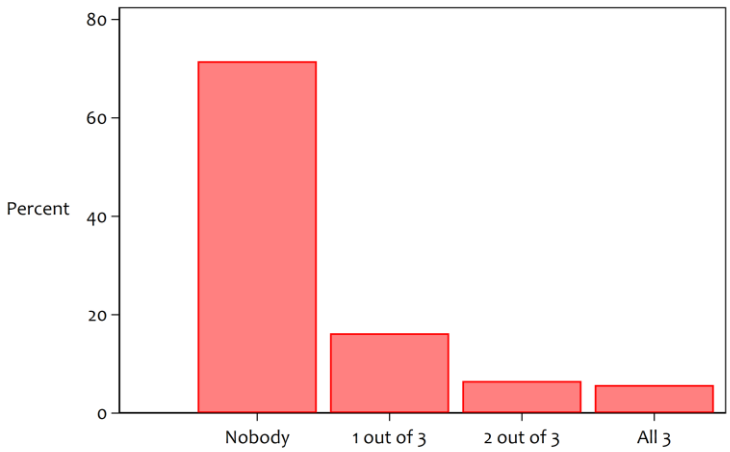


Figure 6.7 shows four panels of the 10-bin belief distribution, which are the 2D beliefs for the [80%, 20%] split. Each panel corresponds to one of the four outcomes in the corresponding 1D belief question for the [80%, 20%] split, which is displayed as the top right panel of Figure 6.4 using blue for displaying the data. The two distributions in the top left panel and the bottom right panel display one similar feature: both distributions have two locally modal beliefs for bin #1 and bin #10. On the other hand, the distributions shown in the top right panel and the bottom left panel are both relatively evenly dispersed. The dispersed distributions suggest relatively *low confidence* in the beliefs assigned to the two events, "1 out of 3 people" and "2 out of 3 people", in the corresponding 1D belief question for the [80%, 20%] split.

Figure 6.7: Descriptive Second-Order Beliefs for [80%, 20%] Split

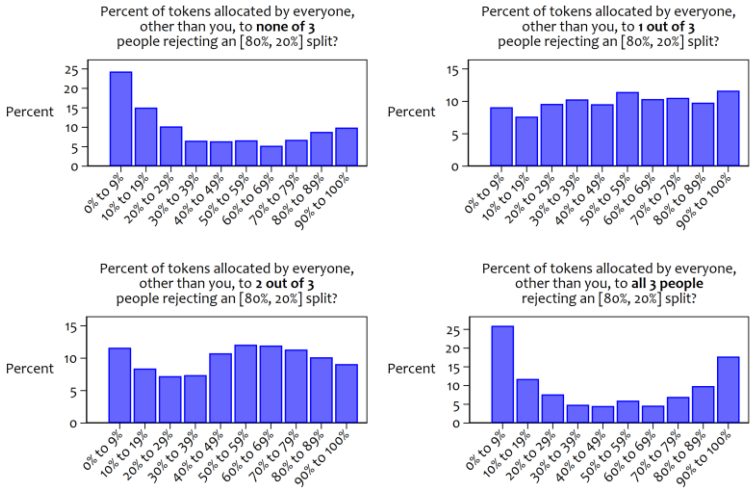
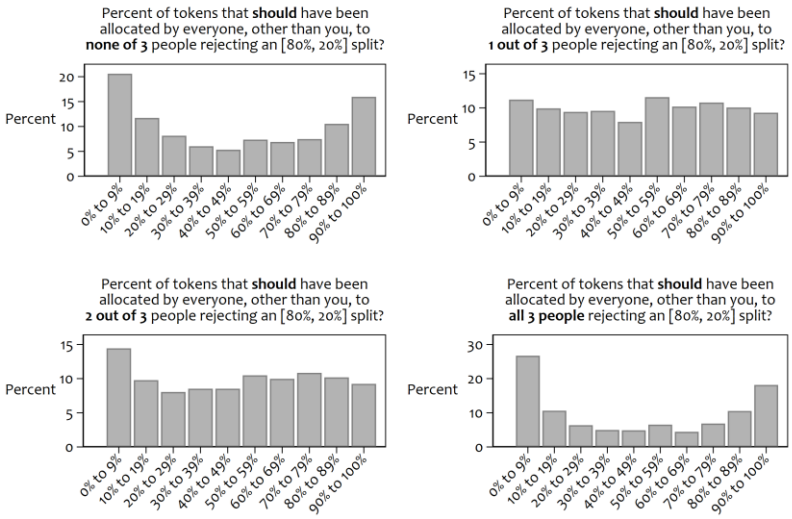


Figure 6.8 shows four panels of the 10-bin distribution, which are the 2N second-order normative beliefs for the [50%, 50%] split. Each of the four panels corresponds to one of the four outcomes in the corresponding 1N belief question for the [50%, 50%] split, which is displayed as the bottom left panel of Figure 6.4 using green for displaying the data.

Figure 6.8: Normative Second-Order Beliefs for [50%, 50%] Split

Figure 6.9 shows four panels of the 10-bin distribution, which are the 2N beliefs for the [80%, 20%] split. Again, each panel corresponds to one of the four outcomes in the 1N belief question for the [80%, 20%] split. The corresponding 1N belief distribution for the [80%, 20%] is the bottom right panel of Figure 6.4, using grey for displaying the data.

Figure 6.9: Normative Second-Order Beliefs for [80%, 20%] Split



6.2 Statistical Models

We now turn to the use of statistical models to formally test the hypotheses of this thesis. One reason for using a statistical model is that there is always some

variation in sample compositions across experimental treatments. For example, if gender affects responses, and there are different mixes across gender from one treatment compared to the baseline, any effect of the treatment could be due to gender and not the treatment. It is possible to check for statistically significant differences in gender, and treat the samples as sufficiently randomised, but those procedures have many inferential issues, particularly with smaller samples.

The primary reason for using statistical models is to be able to characterise the belief distributions in terms of their mean and standard deviation, and then test if the treatments of interest matter for behaviour in a statistically significant and quantitatively significant manner. The latter refers to the effect size: one could have a statistically significant effect that makes no substantive difference in terms of the hypotheses being tested.

There are three aspects of the statistical modelling to be noted. The first is the type of effect size of interest. The second is the type of statistical model appropriate for the first-order beliefs. The third is the type of statistical model appropriate for the second-order beliefs.

6.2.1 Effect Size

Consider the difference between the 1D first-order descriptive beliefs and the 1N first-order normative beliefs for the [80%, 20%] split, displayed in the two right-hand panels of Figure 6.4. Our hypotheses refer to the means of the two distributions, the standard deviations of the two distributions, *and* the mean and standard deviation of the two distributions. The last hypothesis is the one of primary interest, and is a joint hypothesis defined in terms of the first two hypotheses. The first two hypotheses provide insight into why the primary hypothesis is supported, or not, by the evidence.

When a statistical model is estimated to fit the 1D and 1N belief distributions jointly, we will add a binary covariate to capture the effect of the normative posture

of the 1N beliefs in comparison to the descriptive posture of the 1D beliefs. And this covariate, as we see when we discuss the specific statistical models we use below, will capture possible effects on the mean belief *and* possible effects on the standard deviation of beliefs.

Now consider a test of the hypothesis that the covariate affects just the mean belief. There are three types of effect of the covariate that can be tested: the total effect and two types of marginal effect.

The "total effect" shows the effect of just including the covariate for the normative posture, and not allowing for the potential effect of other covariates. Hence if there are other covariates that are correlated with the covariate for the normative posture, the total effect will pick this up. There are some substantive inferential questions for which the total effect is appropriate, and we display it.

The "marginal effect evaluated at the means" shows the effect of the covariate when we allow for the potential effect of other covariates. In statistical models that have a non-linear latent index, such as the ones we use, this marginal effect can and often does differ from the total effect. The "at the means" text refers to the latent index being evaluated at the mean values of the other covariates. If one of those additional covariates is gender, for example, this means that the *average* of the gender covariate would be used to evaluate the marginal effect. This type of marginal effect was used a great deal in the days before cheap processing power, since it simplifies computations that were once costly to undertake.

The "average marginal effect" is the same as the marginal effect evaluated at the means, except that it is evaluated at the actual values of the other covariates, and then those evaluations are averaged. If there are N data points, in our case subjects reporting beliefs, then the marginal effect calculation is undertaken N times, in each case using the actual reported gender for each subject. Then those N marginal effect evaluations are averaged. This method requires N-1 extra computations relative to the marginal effect evaluated at the means, and these are now quite trivial to

undertake. In fact, in economics, and many statistical packages, this type of marginal effect has become the default, to the point where a reference to a "marginal effect" should be assumed to be to the average marginal effect. Since the marginal effect is the one of most interest to us, we also report it along with the total effect.

6.2.2 Statistical Models for the First-Order Beliefs

The data from the 1D and 1N belief elicitations consist of four ordered categories, as displayed in the QSR interfaces in chapter 5 and the bottom axis of Figure 6.4. These are naturally ordered by the number of subjects rejecting a proposal. And they are categories. The discrete nature of the data is sufficient to rule out Ordinary Least Squares, as is the ordered nature of the data. The appropriate statistical model is then an Ordered Probit or an Ordered Logit, and we use the former.

Our interest in statistically characterising the belief distributions is, we have stressed, in terms of the location of beliefs in the distribution (which could be measured by mean, median or mode) as well as the precision of beliefs (measured by the standard deviation). Recall from chapter 4 that we use the term standard *deviation* of beliefs, and sharply distinguish that from the standard *error* of the average of beliefs: see Figure 4.5. This requirement is why we then use an Ordered Probit model that explicitly models the standard deviation of the dependent variable and allows it to vary with covariates: that is, a model that allows for what is called "heteroskedasticity" in econometrics and statistics. In words, we want to allow the standard deviation of beliefs to be different in the 1D belief data to what it is in the 1N belief data, and this is what heteroskedasticity allows. The default assumption in canned statistical packages is, all too often, to model the standard deviation of the dependent variable but to assume "homoskedasticity", the assumption that the variance is the same for all of the data.

Hence, we employ a Heteroskedastic Ordered Probit model for the evaluation

of the 1D and 1N belief responses. These models are quite standard in major statistical packages, such as *Stata*. And they also allow the correct calculation of total effects and (average) marginal effects, which we will utilise.

6.2.3 Statistical Models for the Second-Order Beliefs

The data from the 2D and 2N belief elicitations, displayed in Figure 6.5, consist of 10 categories defining 10 intervals for the token allocations in 1D or 2D, respectively. These intervals refer to percentage token allocations: for example, the second bin is labelled 10% to 19%. The interval-censored nature of the data mean that one should allow for that in the statistical model, and that is now common in major statistical packages using Interval Regression.[33]

One caveat with using Interval Regression is that it defaults to using a Gaussian error term, which assumes that the dependent variable can range between $\pm\infty$, whereas in our data the dependent variable can only range between 0 and 100. One could extend the interval regression model to allow for a Beta error term, which is constrained to lie in the unit interval, and then just model the fraction of token allocations rather than the percent (hence the second bin would reflect data between 0.1 and 0.19, rather than 10% and 19%). This issue is more apparent than real for our data, since we are focusing on the effects of covariates, critically the binary covariate reflecting the use of a normative posture when eliciting beliefs. And the data, particularly when pooled to include the normative beliefs, are not clustered close to 0% or 100%. If we use the standard interval regression estimator, we are able to use the pre-existing software to calculate total effects and (average) marginal effects, which we do.

A more serious caveat with using Interval Regression is that many of the belief distributions do not have the shape of a Normal distribution, even when one ignores

[33] When one uses covariates for the standard deviation of beliefs, as we must, the usual Interval Regression must allow for hetereoskedastic responses as well. The *Stata* implementation of Interval Regression allows for this option.

the tails below 0% or above 100%. This shape is not needed if the covariates of a model explain much of the non-Normal behaviour, since the Normal distribution is used for the *residual* data observed *after* conditioning on these covariates. But if one using the Interval Regression model to characterise unconditional data, or data only conditioned on one covariate such as the normative posture of the question for which beliefs are being elicited, this can be a serious consideration for the data we observe. The Beta Interval Regression model is more flexible in this respect, but a more popular alternative is to again use the Heteroskedastic Ordered Probit model (e.g., Reardon et al., 2017 and Fahle and Reardon, 2018). In this instance, the 10 bin intervals from the belief elicitation can be viewed as 10 categories, compared to the 4 categories for the first-order belief distribution data. And the estimated "cut-points" of the model can then account for the non-Normal distributions as tokens are allocated across these categories.[34] It is also possible to discriminate between the use of an Ordered Probit model and an Interval Regression model, with or without heteroskedasticity, by directly comparing the aggregate log-likelihood values as demonstrated in StataCorp (2021, p. 100). For the data and model used here, the Heteroskedastic Ordered Probit model characterises the belief distributions much better than the Interval Regression model with heteroskedasticity.

6.3 Hypothesis Tests

We can directly test for norm/belief consistency at the First-Order level and then at the Second-Order level for each outcome. Tests for joint consistency over these levels and outcomes require more structure in terms of the econometrics. But tests for joint consistency depend on there being consistency at each level, so it would be a strange econometric magic at work to see different qualitative inferences.

[34] One implication of using the Heteroskedastic Ordered Probit model is that tests of the mean and standard deviation refer to the latent, continous index estimated as part of the model. These are not exactly the same as the mean and standard deviation of the belief distribution defined over the "coarsened" intervals, but are sufficient approximations for our inferential purposes.

For the purpose of this thesis, we will test the hypothesis of belief consistency at each level, i.e., 1D first-order descriptive belief being consistent with the 1N first-order normative belief, and the 2D second-order descriptive belief being consistent with the 2N second-order normative belief.

6.3.1 Hypothesis Tests for First-Order Belief Consistency

The model for testing the *First-Order* beliefs is a Heteroskedastic Ordered Probit model. Covariates include all treatments and a handful of core demographics. The treatments include the normative context of the belief since that is the focus of the primary hypothesis. Other treatments included are the order effect treatment (whether the subject was asked about the [50%, 50%] proposal first or second), exposure to the cartoon treatment for the First-Order normative beliefs, and endowment of the R100 and R300, and a landing page treatment which prompt to encourage consistency. All treatments were applied at random. The demographics are gender, whether the subject reports being Asian ethnicity, whether the subject reports a "very broke" financial condition, and whether the subject reports a "broke" financial condition.

For each of the two proposals, the display in Figure 6.10 shows the significance of the effect of the normative context on the belief distribution. There are four outcomes, and the model recognises that less weight in the normative belief for one outcome has to mean a corresponding greater weight for some or all of the normative belief for all other outcomes as a whole. So, for a given proposal split, a higher (lower) normative belief for some outcome is always offset by a lower (higher) weight for some other outcome or outcomes.

The null hypothesis for the First-Order beliefs is to test that the normative context had *no effect* on the belief distributions, therefore we would expect the 1D descriptive beliefs and the 1N normative beliefs to be the same. However, the distributions displayed in Figures 6.10 and 6.11 have no controls: the two

distributions here are unconditional on any treatment, such as the cartoon treatment, or demographic characteristics such as gender. The sole comparison is between descriptive and normative context, so these displays can be considered as being conditional solely on being *descriptive* or *normative*.

Figure 6.10: Comparing First-Order Beliefs for [50%, 50%] Split

Heteroskedastic Ordered Probit model with controls for treatments and demographics has a *p*-value of 0.001 for the null hypothesis that Normative and Descriptive beliefs are the same (the *p*-value for the mean is less than 0.001, and the *p*-value for the standard deviation is 0.272)

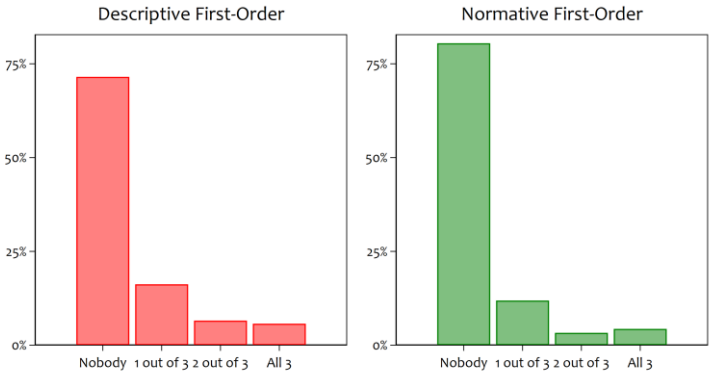


Figure 6.11: Comparing First-Order Beliefs for [80%, 20%] Split

Heteroskedastic Ordered Probit model with controls for treatments and demographics has a *p*-value less than 0.001 for the null hypothesis that Normative and Descriptive beliefs are the same (the *p*-value for the mean is 0.001, and the *p*-value for the standard deviation is less than 0.001)

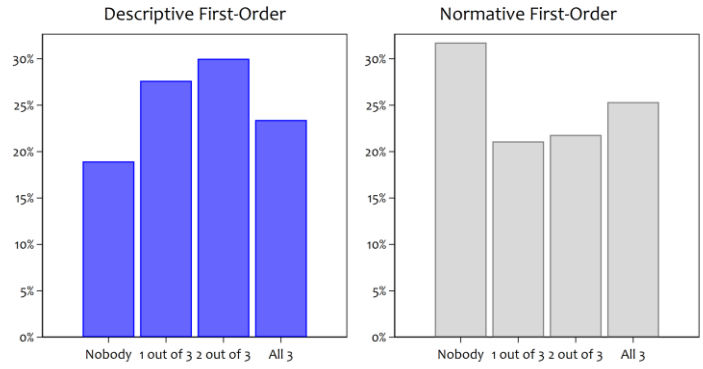


Figure 6.10 compares the First-Order beliefs 1D and 1N for the [50%, 50%] proposal, and are displayed side-by-side. The null hypothesis that both beliefs are the same is rejected with a *p*-value less than 0.001, primarily due to differences in the average beliefs. Normative beliefs put greater weight on nobody rejecting the proposal. Figure 6.11 shows the same tests for the beliefs about the [80%, 20%] proposal. Again, we see the hypothesis of equality of beliefs rejected with a *p*-value less than 0.001, and in this case it is due to differences in both average beliefs and

the dispersion of beliefs. Normative beliefs again put a higher weight on nobody rejecting the proposal, and the size of the difference is much larger than for the [50%, 50%] proposal. The descriptive beliefs are much higher for the two intermediate responses, that only 1 or 2 out of 3 people would reject the proposal. Both of these results clearly reject the null hypothesis that First-Order descriptive and normative beliefs are the same, no matter what the proposal is. These results also point to the dangers of drawing conclusions about belief distributions based on eliciting the modal belief: the modes for the [50, 50%] proposal are the same, but the distributions differ.

Figures 6.12 and 6.13 confirm these results, looking at the total effect and average marginal effect on each possible belief outcome of the normative phrasing of the belief question. We see the significant increase in beliefs about nobody responding, necessarily offset by lower weights on other outcomes.

Figure 6.12: Total Effect of Normative Context
on First-Order Beliefs about Rejection of Proposals

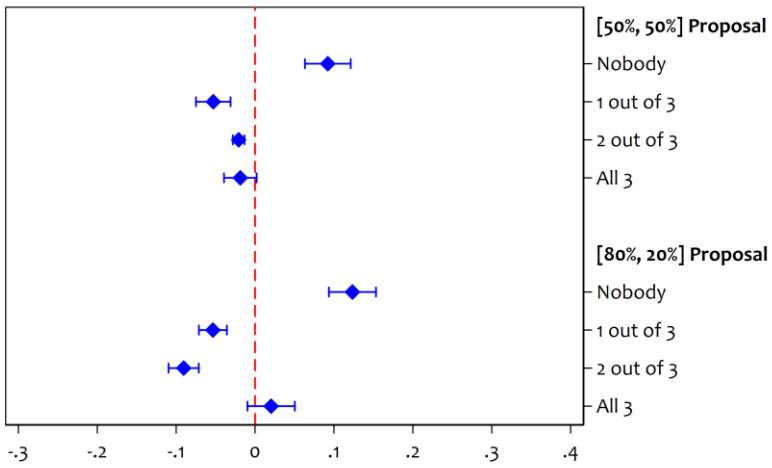
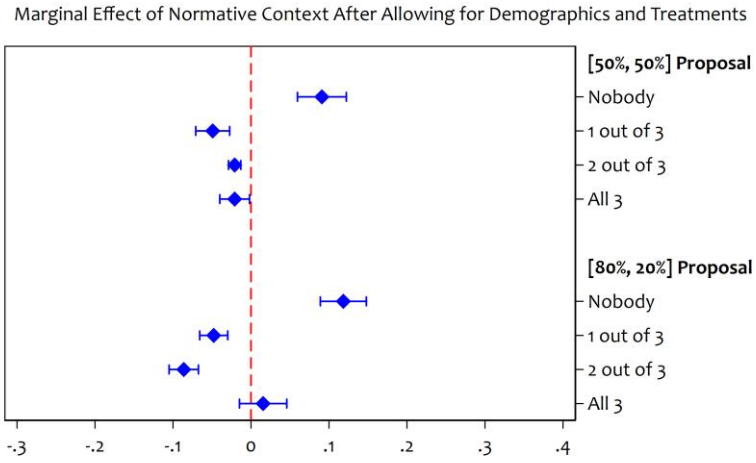


Figure 6.13: Average Marginal Effect of Normative Context on First-Order Beliefs about Rejection of Proposals



6.3.2 Hypothesis Tests for Second-Order Belief Consistency

The model for testing the Second-Order belief distributions is also a Heteroskedastic Ordered Probit model. Covariates again include all treatments and a handful of core demographics. Again, total effects and average marginal effects of the normative context are separately displayed. Norm consistency is first measured at the granular level of each of the possible four outcomes with respect to the First-Order beliefs. The granular outcome level allows direct tests of the null hypothesis.

Figure 6.14 through Figure 17 display the raw belief distributions for the [50%, 50%] split in comparison between the 2D descriptive belief distribution and the 2N normative belief distribution.

Figure 6.14: Comparing Second-Order Beliefs for [50%, 50%] Split:
Beliefs that **None of 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments and demographics has a *p*-value less than 0.001 for the null hypothesis that Normative and Descriptive beliefs are the same (the *p*-value for the mean is 0.134 and for the standard deviation is less than 0.001)

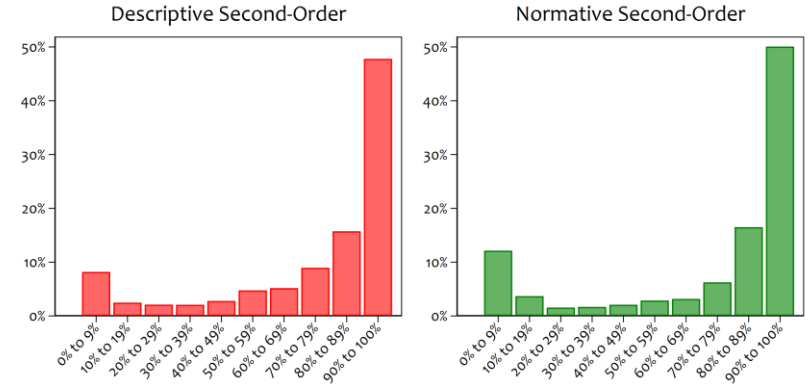


Figure 6.14 begins the evaluation of Second-Order beliefs 2D and 2N, in this case for the [50%, 50%] proposal and the First-Order beliefs attached to the outcome that nobody would reject the proposal. In evaluating these Second-Order beliefs, it is useful to recall the First-Order belief distribution. In this instance, from Figure 6.10 we observe that most of the weight of beliefs was on this outcome. Hence it matters more for our overall inferences: if virtually nobody holds beliefs about some First-Order descriptive *and* normative outcome, then there is much less interest in whether descriptive and normative second-order beliefs about it are the same. In Figure 6.15 the displays have a similar "look to the eye", but they are different overall, primarily due to their dispersion. Normative beliefs allocate more weight to the lowest chances of observing this outcome (viz., between 0% and 9% chance, and between 10% and 19% chance), and less weight to some of the higher chances of observing this outcome (e.g., 70% to 79% chance). Moreover, the displays are intended to guide intuition, but do not control for treatments or demographics, and the statistical model does, to ensure that the hypothesis tests focus solely on the difference between descriptive and normative beliefs. The overall hypothesis of equality of the belief distributions is rejected at a *p*-level less than 0.001.

Figure 6.15: Comparing Second-Order Beliefs for [50%, 50%] Split:
Beliefs that **1 of 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value of 0.008 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is 0.002 and for the standard deviation is 0.824)

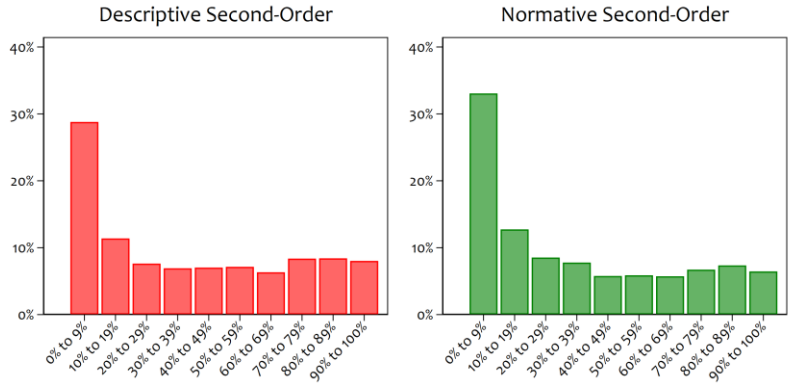


Figure 6.15 then moves to the next outcome, again for the [50%, 50%] proposal, the fraction of beliefs that only 1 of 3 people would reject the proposal. Here we find that the averages are quite different, although we cannot reject the hypothesis that the dispersion of beliefs is the same. Overall, we reject the hypothesis that the two distributions are the same with a *p*-value of 0.0008, primarily due to the greater weight on the lowest outcome (0% to 9%) for normative beliefs.

Figure 6.16: Comparing Second-Order Beliefs for [50%, 50%] Split:
Beliefs that **2 of 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value of 0.122 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is 0.043 and for the standard deviation is 0.589)

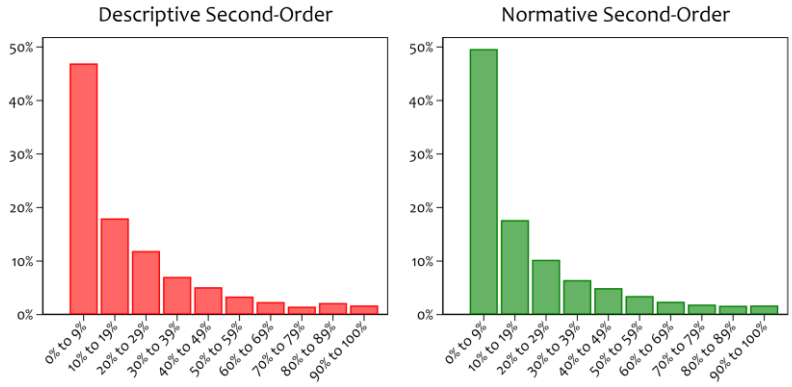
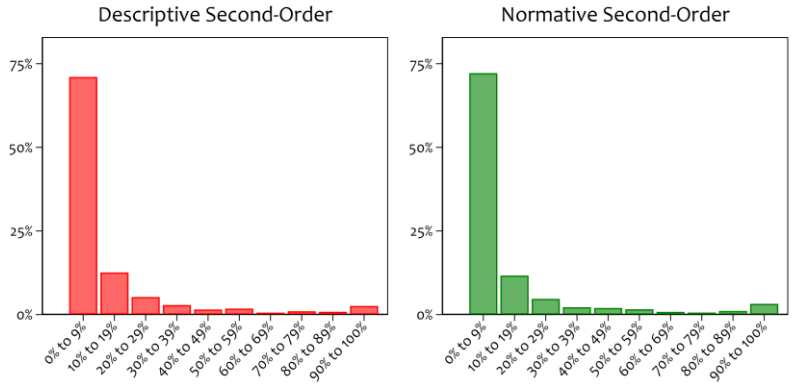


Figure 6.16 then considers beliefs over the outcome of 2 of 3 people rejecting the [50%, 50%] proposal, and here we have the first instance where we cannot reject the null hypothesis of consistent Second-Order beliefs. The *p*-value for the overall

test of the hypothesis is only 0.122, although there is some evidence of a significant difference in means. On the other hand, Figure 6.10 shows us that very few subjects held First-Order beliefs that this outcome would occur. So, we have second-order consistency of beliefs about First-Order beliefs that virtually nobody puts any weight on.

Figure 6.17: Comparing Second-Order Beliefs for [50%, 50%] Split:
Beliefs that **All 3** People Rejected the Proposal
Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value of 0.134 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is 0.051 and for the standard deviation is 0.097)



The same qualitative result is obtained for the final outcome of the Second-Order beliefs for the [50%, 50%] proposal, that 3 out of 3 people would reject the proposal. The overall hypothesis test for the distribution only has a *p*-value of 0.134, despite there being some evidence of a different mean and standard deviation. But again, this is an outcome with very little weight from the First-Order beliefs, so again we have second-order belief consistency over empirically irrelevant First-Order beliefs.

The overall conclusion about the Second-Order beliefs for the [50%, 50%] proposal is a rejection of the hypothesis of consistency of normative and descriptive beliefs. For the two First-Order outcomes of any empirical interest, the significance levels are below 0.001, and second-order belief consistency is only inferred for First-Order beliefs of virtually no empirical interest. Since belief consistency over all outcomes is the weighted product of belief consistency over all four, these results imply a rejection of the null hypothesis for the Second-Order beliefs for the [50%,

50%] proposal. This result for this joint hypothesis test is the same, but particularly striking, if we account for the greater weight on the first two outcomes.

Figures 6.18 through 6.21 display comparable results of hypothesis tests for the [80%, 20%] split. Here we must pay attention to belief consistency for all four possible outcomes, since we recall from Figure 6.11 that both descriptive and normative beliefs assigned considerable weight to each outcome.

Figure 6.18: Comparing Second-Order Beliefs for [80%, 20%] Split:
Beliefs that **None of 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value less than 0.001 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is less than 0.001 and for the standard deviation is 0.011)

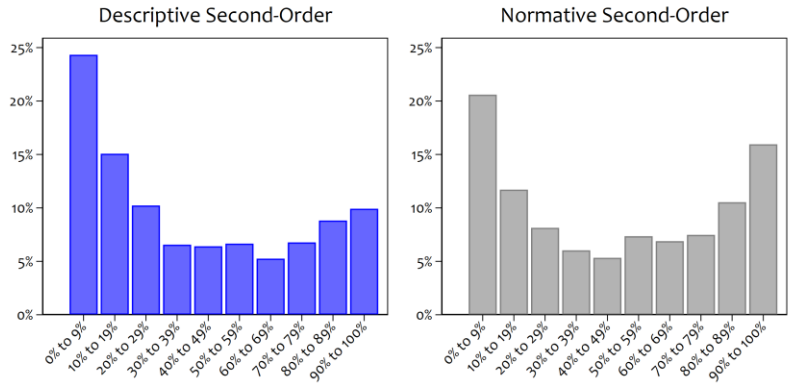


Figure 6.19: Comparing Second-Order Beliefs for [80%, 20%] Split:
Beliefs that **1 of 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value of 0.225 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is 0.085 and for the standard deviation is 0.769)

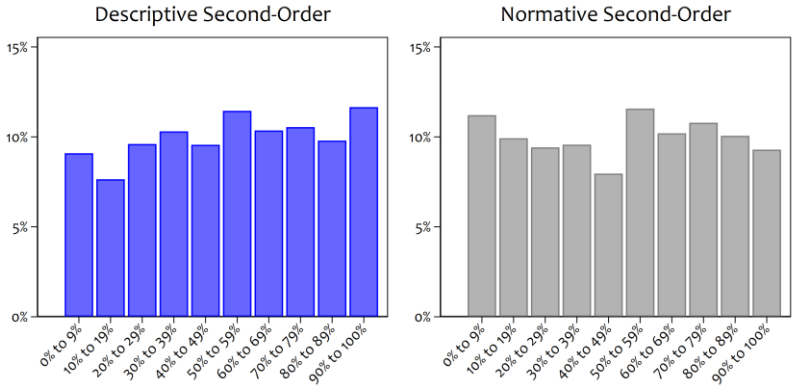


Figure 6.20: Comparing Second-Order Beliefs for [80%, 20%] Split:
Beliefs that **2 of 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value of 0.021 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is 0.066 and for the standard deviation is 0.025)

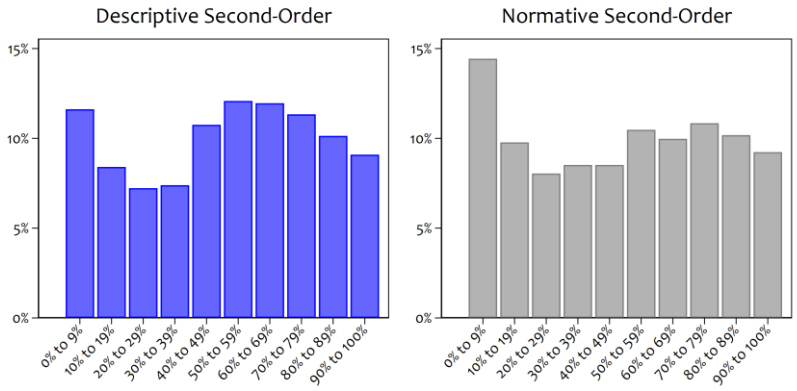
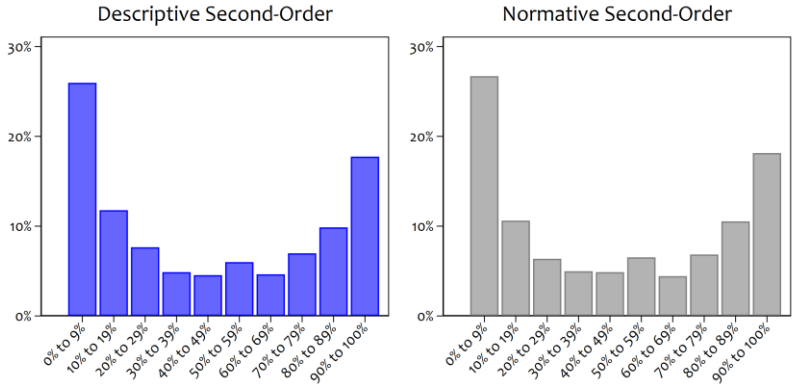


Figure 6.21: Comparing Second-Order Beliefs for [80%, 20%] Split:
Beliefs that **All 3** People Rejected the Proposal

Heterokedastic Ordered Probit model with controls for treatments
and demographics has a *p*-value of 0.656 for the null
hypothesis that Normative and Descriptive beliefs are the same
(the *p*-value for the mean is 0.840 and for the standard deviation is 0.400)



We find from Figure 6.18 and Figure 6.20 that we can reject belief consistency of Second-Order beliefs for beliefs about the outcome that nobody would reject the [80%, 20%] proposal and beliefs about the outcome that 2 out of 3 people would reject the [80%, 20%] proposal, respectively. The *p*-values for these hypothesis tests are less than 0.001 and 0.021, respectively. For the other two outcomes, in Figure 6.19 and Figure 6.21, we cannot reject the hypothesis of second-order belief consistency: the *p*-value for beliefs about the outcome that 1 out of 3 people would reject the proposal is 0.22, and the *p*-value for beliefs about the outcome that all 3 people would reject the proposal. Figures 6.22 through 6.29 provide a rich

quantification of the effects of the normative belief question on each of the ten possible responses to the second-order questions.

Figure 6.22: Total Effect of Normative Context
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **None of 3** People Rejected the Proposal

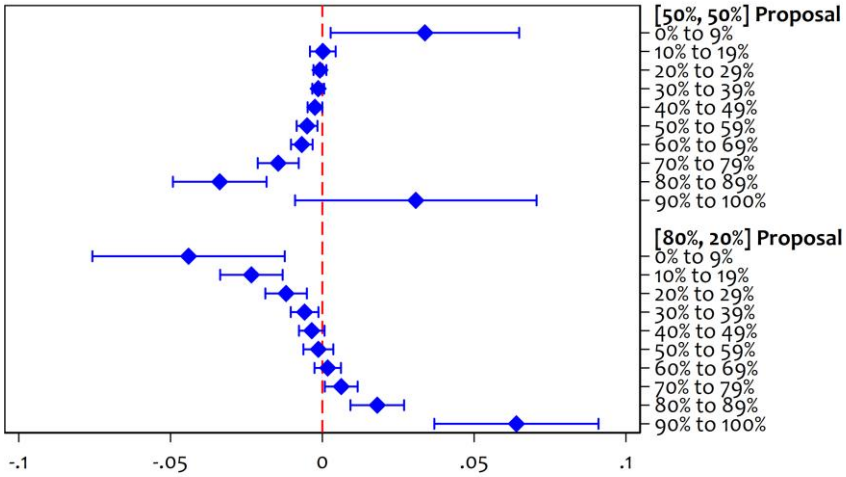


Figure 6.23: Total Effect of Normative Context
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **1 of 3** People Rejected the Proposal

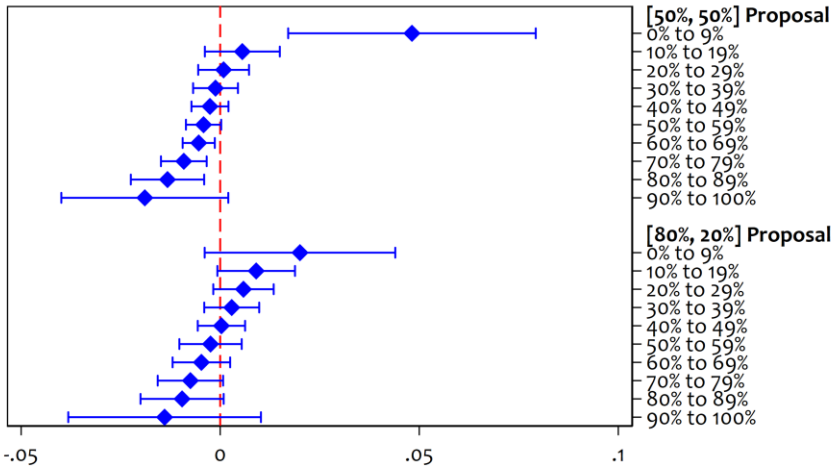


Figure 6.24: Total Effect of Normative Context
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **2 of 3** People Rejected the Proposal

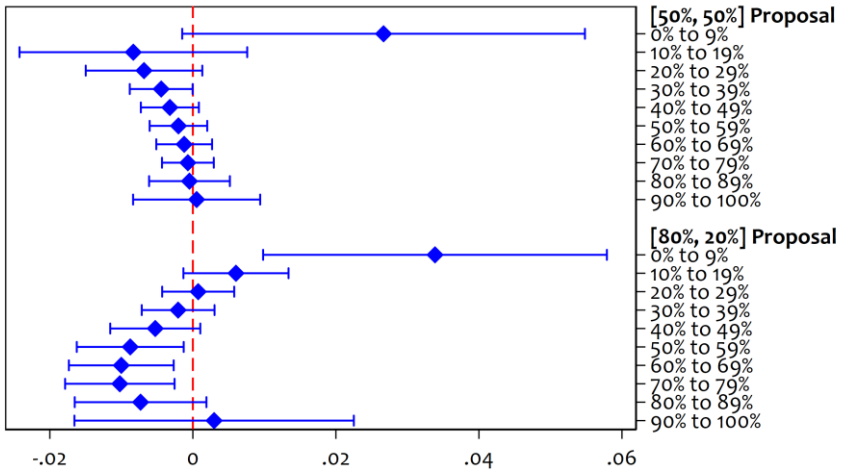


Figure 6.25: Total Effect of Normative Context
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **All 3** People Rejected the Proposal

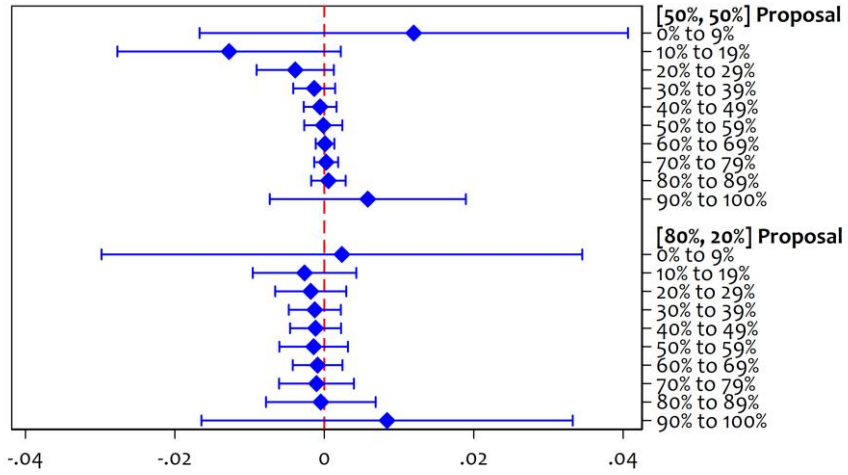


Figure 6.26: Average Marginal Effect of Normative Context on Second-Order Beliefs about Rejection of Proposal: Beliefs that **None of 3** People Rejected the Proposal

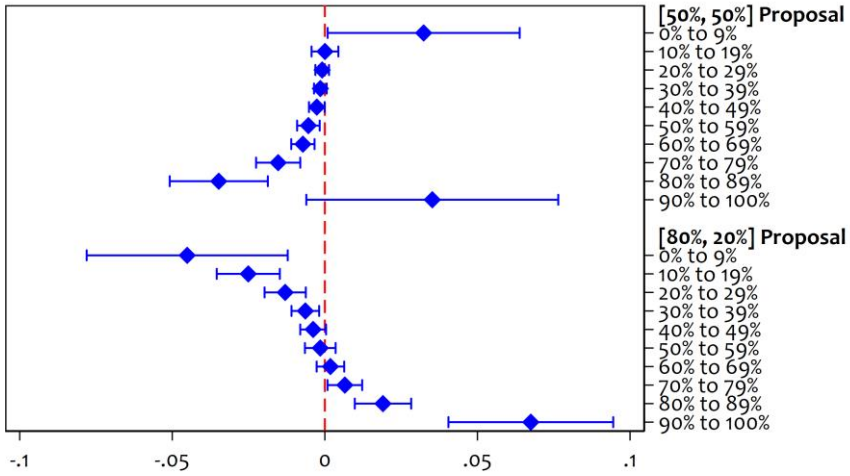


Figure 6.27: Average Marginal Effect of Normative Context on Second-Order Beliefs about Rejection of Proposal: Beliefs that **1 of 3** People Rejected the Proposal

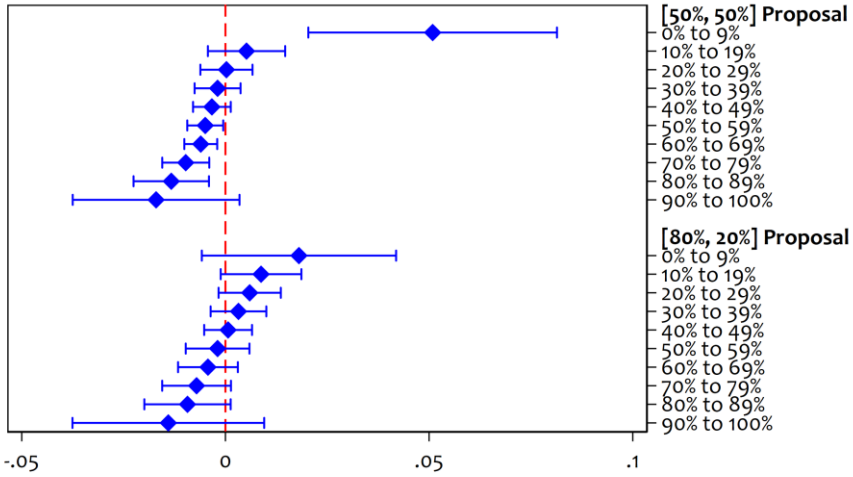


Figure 6.28: Average Marginal Effect of Normative Context on Second-Order Beliefs about Rejection of Proposal: Beliefs that **2 of 3** People Rejected the Proposal

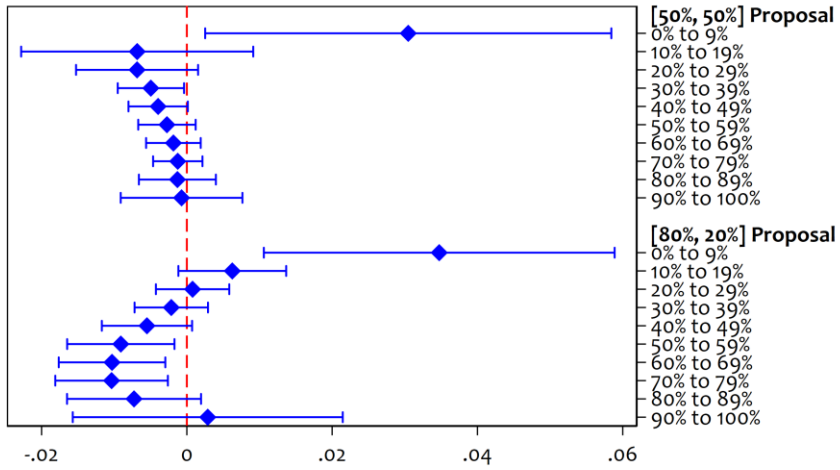
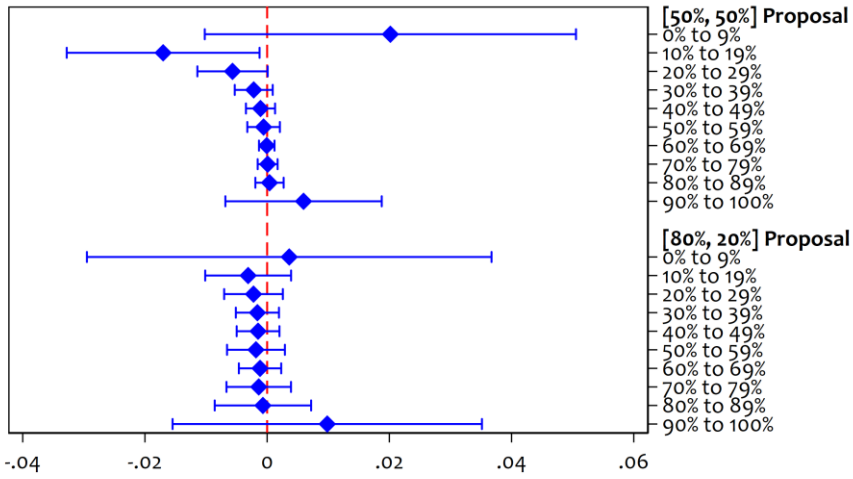


Figure 6.29: Average Marginal Effect of Normative Context on Second-Order Beliefs about Rejection of Proposal: Beliefs that **All 3** People Rejected the Proposal



The overall conclusion about the Second-Order beliefs for the [80%, 20%] proposal is a rejection of the hypothesis of consistency of normative and descriptive beliefs for two of the outcomes over which First-Order beliefs were elicited, but support for consistency for the other two outcomes. Unlike the First-Order belief data for the [50%, 50%] proposal, in this case all four possible outcomes matter empirically. Since belief consistency over all outcomes is the weighted product of belief consistency over all four, and all four have comparable weights, these results

imply a rejection of the null hypothesis for the Second-Order beliefs for the [80%, 20%] proposal.

6.3.3 Cartoon Treatment and other Treatments

There are four treatments in this study, but there is no significant effect on the belief data in any treatment. For demonstration purpose, I consider the statistical results from the cartoon treatment. Figure 6.30 displays the results of the total marginal effects of the cartoon treatment for the First-Order beliefs about rejection of the two proposal splits.

Figure 6.30: Total Effect of Cartoon Treatment
on First-Order Beliefs about Rejection of Proposals

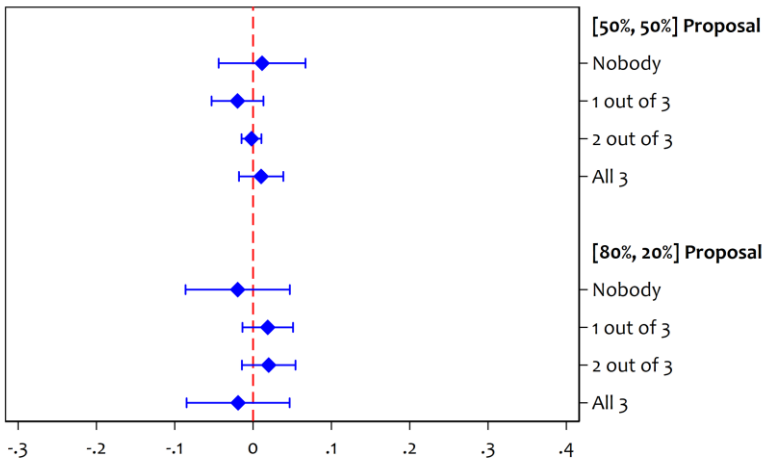


Figure 6.31 displays the average marginal effects of the cartoon treatment for the First-Order beliefs about rejection of the two proposal splits in the normative context.

Figure 6.31: Average Marginal Effect of Cartoon Treatment on First-Order Beliefs about Rejection of Proposals

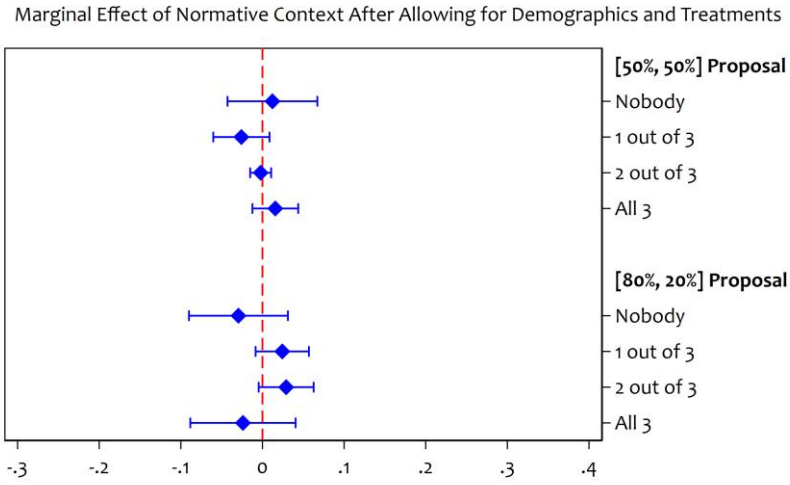


Figure 6.32, Figure 6.33, Figure 6.34, and Figure 6.35 display the results of the total effects of the cartoon treatment for the Second-Order beliefs about rejection of the two proposal splits.

Figure 6.32: Total Effect of Cartoon Treatment on Second-Order Beliefs about Rejection of Proposal: Beliefs that **None of 3** People Rejected the Proposal

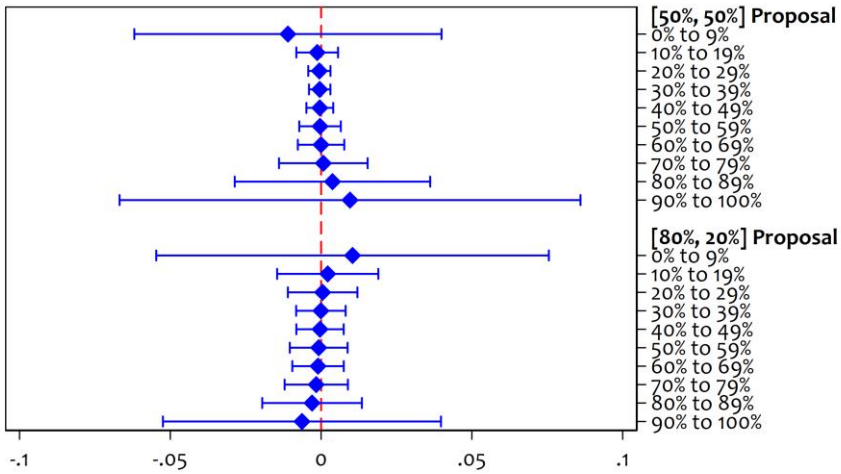


Figure 6.33: Total Effect of Cartoon Treatment
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **1 of 3** People Rejected the Proposal

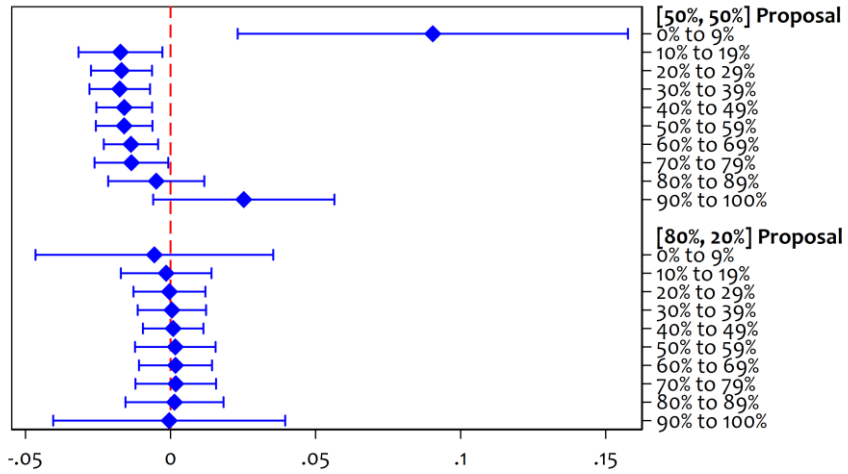


Figure 6.34: Total Effect of Cartoon Treatment
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **2 of 3** People Rejected the Proposal

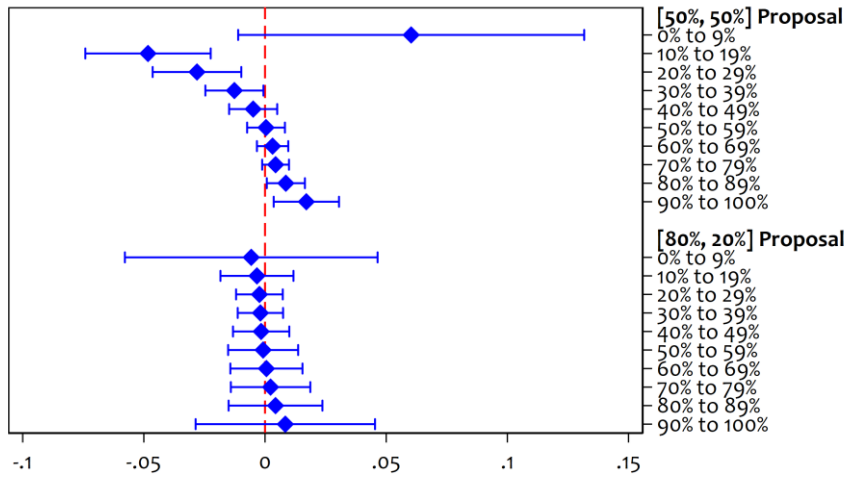


Figure 6.35: Total Effect of Cartoon Treatment
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **All 3** People Rejected the Proposal

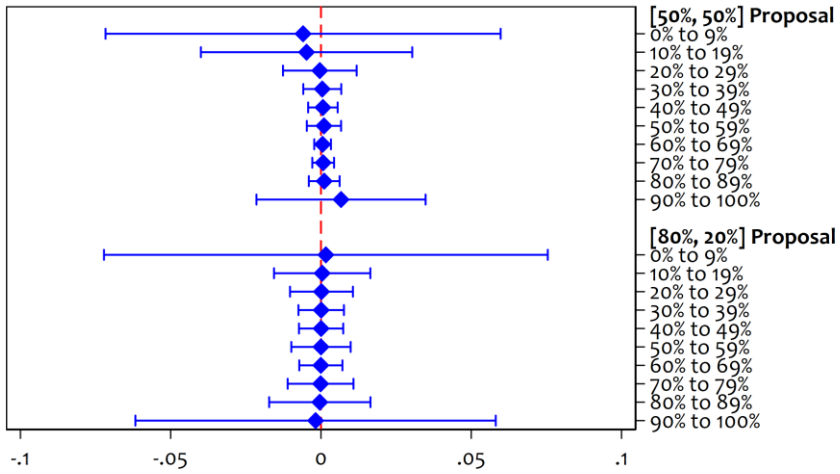


Figure 6.36, Figure 6.37, Figure 6.38 and Figure 6.39 displays the average marginal effects of the cartoon treatment for the Second-Order beliefs about rejection of the two proposal splits in the normative context.

Figure 6.36: Average Marginal Effect of Cartoon Treatment
on Second-Order Beliefs about Rejection of Proposal:
Beliefs that **None of 3** People Rejected the Proposal

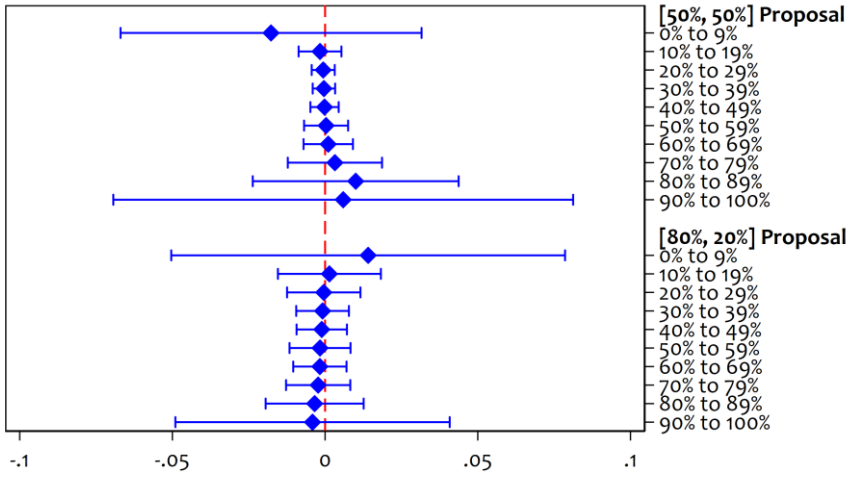


Figure 6.37: Average Marginal Effect of Cartoon Treatment on Second-Order Beliefs about Rejection of Proposal: Beliefs that **1 of 3** People Rejected the Proposal

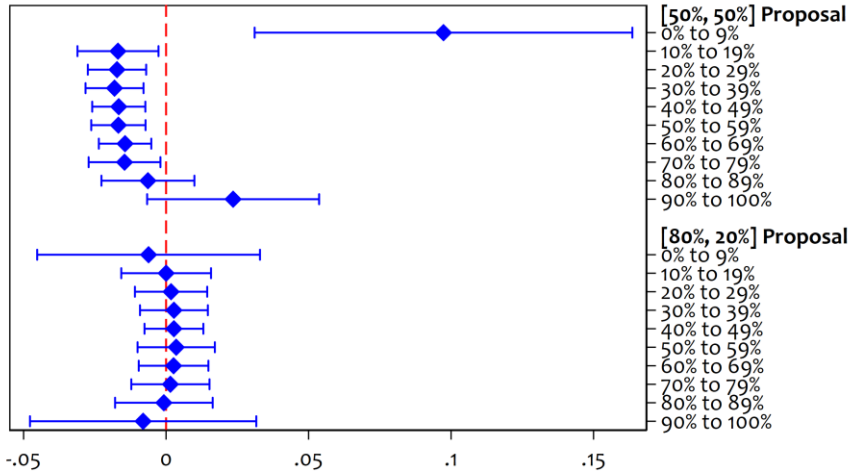


Figure 6.38: Average Marginal Effect of Cartoon Treatment on Second-Order Beliefs about Rejection of Proposal: Beliefs that **2 of 3** People Rejected the Proposal

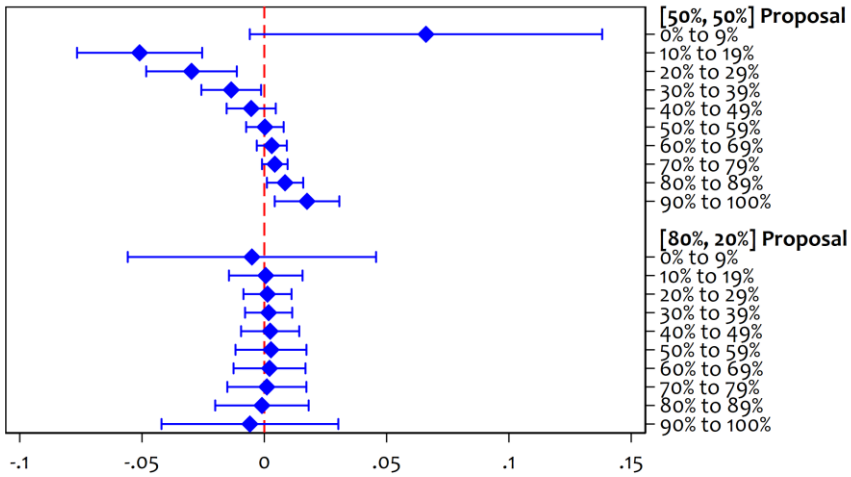
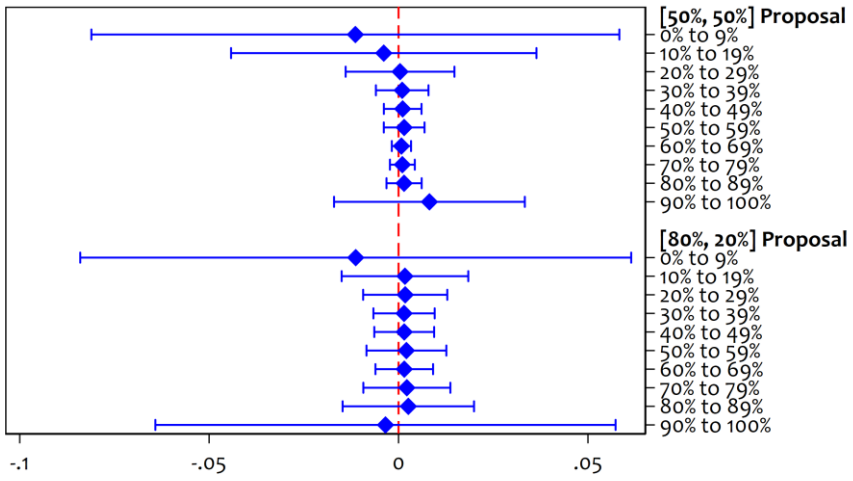


Figure 6.39: Average Marginal Effect of Cartoon Treatment on Second-Order Beliefs about Rejection of Proposal: Beliefs that **All 3** People Rejected the Proposal



There is no effect on First-Order beliefs, whether the proposal was [50%, 50%] or [80%, 20%]: see Figures 6.30 and 6.31. For Second-Order beliefs there are no effects for the First-Order beliefs about the "None of 3" and "All 3" outcomes, whether the proposal was [50%, 50%] or [80%, 20%]. There was some significant effect on Second-Order beliefs for the First-Order beliefs about the "1 out of 3" and "2 out of 3" outcomes, but only when the proposal was [50%, 50%]. Of course, both of these outcomes received very little weight in the beliefs over all four outcomes, as shown in Figure 6.10. We conclude that the Cartoon treatment did not have a significant effect on beliefs.

6.3.4 Normative Reference Networks

One way to judge if the sample constitutes a Normative Reference Network is to judge if the sample reports "consensus" about support for a series of value statements. In chapter 5 the measure of consensus derived and explained by Tastle and Wierman (2007) was proposed to make this judgment. A score of 1 indicates perfect consensus of a sample, and 0 indicates perfect complete lack of consensus. The average score, over the 16 scores for each statement, was 0.59 for the 255 subjects in the UG task, 0.56 for the 396 subjects in the Beliefs task, and 0.57 pooled

over both tasks. The lowest score for the pooled responses was 0.48 for the **Abortion** question, and the highest was 0.70 for the **ServTax** question. These results provide some support for the claim that our subjects constitute a Normative Reference Network for our purposes. There were no major differences in scores when comparing the sub-samples for each task.

6.4 Conclusions

We find a significant difference in rejection rates for the [50%, 50%] and [80%, 20%] proposals, with very few subjects rejecting the former and about a quarter of subjects rejecting the latter.

First-Order beliefs are dramatically different for the two proposals. They are centered on nobody rejecting the [50%, 50%] proposal, but all of the possible outcomes receive support for the [80%, 20%] proposal. This characterisation generally applies to both descriptive and normative beliefs.

We find statistically significant evidence for differences in descriptive and normative beliefs for each of the proposals. The evidence for a difference is most prominent for the response that nobody will reject the proposal in the case of the [50%, 50%] proposal, and of course that outcome is the most important empirically for that proposal. The evidence for a difference spans several outcomes in the case of the [80%, 20%] proposal, but again is greatest for the response that nobody will reject the proposal.

The Second-Order beliefs for the [50%, 50%] proposal leads to a rejection of the hypothesis of consistency of normative and descriptive beliefs. This rejection comes from the two First-Order outcomes of any empirical interest. The Second-Order beliefs for the [80%, 20%] proposal leads to a rejection of the hypothesis of consistency of normative and descriptive beliefs for two of the outcomes over which First-Order beliefs were elicited, but support for consistency for the other two outcomes. Belief consistency over all outcomes is the weighted product of belief

consistency over all four possible outcomes, so we reject the null hypothesis for the Second-Order beliefs for the [80%, 20%] proposal.

We find neither the cartoon treatment nor any other treatment (Endowment Effect treatment, Landing page treatment, and Order Effect treatment) had any significant effect on First-Order or Second-Order beliefs.

Chapter 7 - Conclusions

7.1 Summary and Theoretical Extensions

This thesis started with a discussion in Chapter 1 motivating choice of the social structure account of social norms developed by Bicchieri as the preferred theoretical approach to understanding social norms. Chapter 2 then identified and reviewed three kinds of conditions for norm identification suggested by this theory, and identified empirical hypotheses corresponding to each condition. After the review of Bicchieri's conceptual analysis, Chapter 3 provided some suggestions on the philosophical framework for interpreting the intentional concepts applied in Bicchieri's analysis of social norms.

The proposed philosophical framework aligns the empirical approach with the literature in economics on elicitation of revealed preferences and incentivised beliefs. Chapter 4 then provided critical reviews of some key experimental contributions by Bicchieri and co-authors applying her theory. The chapter also provided a critical review of a widely used toolbox in current economic literature for norm elicitation. This leads us to the experimental protocols designed for the thesis. These are specified in Chapter 5, along with the procedures actually implemented online at the University of Cape Town, South Africa, in November 2021. Statistical modelling of the results of the experiment was presented in Chapter 6.

This thesis encourages some theoretical reflections on Bicchieri's theory of social norms. First, the research presented here motivates various conceptual clarifications of the theoretical work. Second, attention to experimental design considerations assesses the operationalisability of the theory. The techniques applied to elicit and estimate beliefs move beyond the limitation of point estimates which, discard important information. The techniques used in the thesis allow for the recovery of subjects' entire underlying subjective belief distributions, while inducing risk neutral responses. This allows for measurement of belief consistency with respect to both bias and confidence.

Using controlled laboratory experiments with real rewards and incentivised elicitation of beliefs, in addition to proper econometrics for the data analysis, we find that the hypothesis generated by Bicchieri's conceptual analysis of social norms does not hold. We conclude that Bicchieri's theory incorporates an incomplete specification of the idea of belief consistency.

Several extensions follow naturally from the work in the thesis.

On the theoretical side, one extension would be to investigate further the "scope issue" pointed out in section 2.4. Excluding moral norms on the grounds that these are unconditional, as Bicchieri suggests, is not only inconsistent with standard literature in moral psychology, but also it severely restricts the scope of her analysis. On one hand, it is highly doubtful whether many philosophers would agree with Bicchieri's view of moral norms as generally unconditional. On the other hand, it is also highly unlikely that such understanding can be supported by empirical findings. Theoretical work to allow for the conditionality of moralised norms can expand the scope of Bicchieri's analysis in a way that is consistent with her ambitions with respect to empirical social theory and policy applications.

A second extension is to make more of the concept of reference network applied by Bicchieri for norm identification. As reviewed in this thesis, reference network alignment is one of the conditions suggested by Biccchieri for norm identification. If Bicchieri's conjecture that equilibria of expectations should hold within a reference network is correct, then an extension of this argument as applied to norm nudging would shift the focus from changing individuals' beliefs to measures that might better promote reference network alignment as the primary target.

A major problem for Bicchieri's theory is that if it insistently relies on the distinction between conditionality and unconditionality, but then identifies moral norms as socially important unconditional beliefs, then it means that this theory will not apply to empirically prevalent social phenomena such as those modelled by

Kuran (1995). Excluding moral norms from the scope of the theory renders it much less practically important than it seems to be.

I will illustrate this by reference to a hypothetical but realistic example. On 22 May 2015, the results of the Marriage Equality Referendum made Ireland the first country in the world to legalise same-sex marriage via a popular vote, even though Ireland was a country, by then, where 84% of Irish people still identified as Catholic (Tobin 2016). Let us imagine a hypothetical Irish voter, Aoife. In 2014 when public opinion in Ireland was still against same-sex marriage, when Aoife was publicly asked to announce her opinions at an Irish bar about same-sex marriage, she acted as if she was against same-sex marriage, that is in line with public opinion. Whereas in the same year, when Aoife attended a cocktail party at Harvard University where public opinion is known to be progressive, Aoife revealed her private opinion as supporting same-sex marriage.

This is a case of what Kuran (1995) models as "preference falsification". Let's consider how it might be accommodated by Bicchieri's model. As reviewed in Chapters 2 and 4, empirical application of Bicchieri's theory relies on belief consistency. Bicchieri (2017) argues that "when second-order beliefs [normative] are mutually consistent but systematically inaccurate, we know that people uphold a norm they dislike. If their doubts are not shared, the norm will persist" (p. 74). The statement by Bicchieri suggests that when we uncover perceived across-individual belief consistency of second-order normative (2N) beliefs in a society with large variance (inaccuracy of beliefs) among individuals, it indicates that in fact some of the perceived 2N beliefs might be outcomes of falsified first-order normative (1N) beliefs. This means, according to Bicchieri, that where a Kuran case is detected, it must result from individuals' belief *inconsistency* between the 1N and 2N beliefs of the falsifiers.[35] However, this seems implausible.

In the case of Aoife falsifying her private preference (in Kuran's language) in

[35] Keep in mind that the falsifiers falsify their 1N beliefs.

the Irish bar, it is unclear what Aoife's 1N belief actually is. Recall that in Bicchieri's concept, the 1N belief is belief about what others should do. Aoife's personal 1N belief is that she *herself* should support same-sex marriage (given the fact that she voted for same-sex marriage). She may wish other Irish people to endorse her personal belief, and then her 1N belief (about others) would be "Aoife thinks others should support same-sex marriage". Alternatively, she may think other Irish people should be sincere, and that she may think if other Irish people were against same-sex marriage, they should say so. It seems it is not clear what Aoife's 1N belief in Bicchieri's structure actually is. What about Aoife's 2N belief? Would Aoife expect other Irish people to think that she and others should support same-sex marriage which is a normative belief they do not seem to have (from their publicly endorsed opinion)? Or, would Aoife expect other Irish people to think that she and others should not support same-sex marriage? As it is not clear what Aoife's 1N and 2N beliefs are in this case, it is certainly not straightforward whether there is a conflict between Aoife's 1N and 2N belief.

Thus it is not obvious how Bicchieri's concepts of 1N and 2N beliefs can work here to think about Kuran cases.

Norms about issues such as same-sex marriage tend to be generally moralised. Bicchieri would therefore expect beliefs here to be unconditional. She might therefore say that the case falls outside the scope of her account. This would apply generally to the kinds of cases modelled by Kuran (1995). But then this is inconsistent with Bicchieri's (2017) application of her analysis of norms to reduction of open defecation and female genital mutilation, both of which are widely moralised and about which there is certainly substantial preference falsification in countries where these behaviours are challenged by reformers.

What lies at the foundation of the misalignment between Bicchieri's and Kuran's frameworks then is Bicchieri's regarding morality as unconditional, whereas social norms are conditional. She remains explicit about this in most recent accounts.

However, subsequent to her original presentation of her theory in Bicchieri (2006), she has suggested that excluding moral norms from the scope of her theory is not a problem because personal moral norms are typically cheap talk. Bicchieri (2014) says that “when personal beliefs and normative expectations disagree, I predict that normative expectations, not personal normative beliefs, will guide behaviour. This is in line with what social psychologists have observed: beliefs that are perceived to be shared by a relevant group will affect action, whereas personal normative beliefs often fail to do so, especially when they deviate from socially held beliefs.” (2014, p. 226). This cynical view of standard expressions of moralised preference is endorsed by almost no academic literature outside of old-fashioned Marxist accounts.

Kuran analyses situations by which agents persistently follow a prevailing norm that they privately dislike and would be better off if everyone dropped. However, as long as this doesn’t happen, they are better off following the norm. Most of these social phenomena involve norms which societies regard as morally significant.

The social phenomena that concern both Bicchieri and Kuran are similar: culturally evolved expectations that prescribe and proscribe individuals’ behaviour. They involve human activities which are assessed not in isolation but in connection with the welfare of others. Take the caste norm in India as an example. Kuran stresses the *interdependencies* between different castes that generated pressures to keep individuals loyal to the system. As Kuran analyses, “because society will generally ostracise anyone who abandons the caste system, the potential member of an anticaste colony is likely to withhold his participation until it appears likely to succeed” (p. 134). Even the expression of discontent and resentment against the caste system would be sanctioned. Naturally, as Kuran points out, “because of the costs of sincerity, people conceal their willingness to subvert the system, except perhaps from family members and trusted friends" (p. 134).

An example of a similar case discussed by Bicchieri (2017) is normative tolerance of punishment of children. “A mother may choose” Bicchieri writes, “to overtly beat her child because all the other parents around her do so, and she fears being looked down upon or reprimanded if she does not hit him hard. In fact, she might not like to punish so harshly, but what would the neighbors think of her?” (p. 8). The concept of interdependency is also the key characteristic in this example. What identifies this practice as a social norm in Bicchieri’s model is the feature of interdependency and interconnection with others for some people who choose to abide by the rule.

In addition to concerns with similar social problems, Bicchieri’s and Kuran’s models also share similarity in analytical approaches. First, both theories focus on *conditionality* of norm compliant behaviour. In Bicchieri’s model, it is the alignment between different kinds of social beliefs, whereas in Kuran’s model it is the tradeoffs between three types of utility as determinants of individuals’ preferences to be revealed. Second, both theories couch their models in terms of applications of mentalistic concepts of belief and preference. Third, both models are frameworks grounded in individual choices but reject both completely individualistic or completely structuralist approaches in modelling norms. Therefore, both theories also serve to integrate the individualist and structuralist traditions and show that these approaches are not at all incompatible.

Norm compliant behaviours are modelled by Kuran as preference falsification which refers to “the act of misrepresenting one’s genuine wants, under perceived social pressures” (p. 3). Kuran distinguishes between public preferences and private preferences. Public preferences refer to the preferences individuals convey to others, and private preferences are the preferences they would express in the absence of social pressure. The genuine wants are the ‘private preferences’ and the publicly signaled wants are the ‘public preferences’ in Kuran’s language. Kuran argues that in practice, the publicity of a preference varies along a continuous spectrum: at one

extreme lie one's secret preferences and at the other is advocacy for general distribution. We can imagine intermediate cases on the spectrum, e.g., preferences expressed only to one's partner, or preferences expressed only in academic seminars. We understand at which point along the spectrum an agent reveals her preferences by reference to a tradeoff. The terms of the tradeoff are determined by the agent's utility function as modelled by Kuran (1995).

Kuran introduces three kinds of additive utility (for simplicity of the modelling) which factor in the calculation of the tradeoffs. These are intrinsic utility, reputational utility, and expressive utility. The total utility of an individual i is determined by the tradeoffs which is given by:

$$U' = P^i\big(d(y')\big) + R^i(y') + A^i(x', y'),^{36}$$

in which U^i denotes the total utility of an individual i, while P^i, R^i, and A^i represent intrinsic utility, reputational utility, and expressive utility respectively. Notation d represents the society's decision which depends on all public preferences, including that of individual i's.

Intrinsic utility refers to the utility an agent would gain regardless of what public preferences she endorses in the society in question. It is what 'society's decision' intrinsically brings to individuals (p. 36). As Kuran's points out, for typical members of societies, their influence over outcomes of public choices is generally effectively zero. Thus, common people's public preferences don't affect the intrinsic utility they gain or lose from implemented policies. A counterexample which represents a rare case is that for powerful people like a prime minister, the weight of her public opinion is significant for society's decisions. But for typical agents, intrinsic utility is irrelevant to their choices over which preferences to reveal.

Reputational utility refers to the net payoff a person receives from the various responses to a *public* preference she manifests, hence the reputation an agent gains

[36] This utility function is cited from Kuran (1990, p. 11).

from revealing a particular preference in public. Expressive utility refers to the gains from expressing a private preference in public, given that an agent cares about the alignment between her public opinion and private opinion. Expressive utility reflects the "need for self-assertion" (p. 31), which is to do with the value to the individual of her sense of individuality, autonomy, dignity, and integrity. Such individuality is cultivated maximally by supporting publicly whatever option in a policy choice set the person likes best in private.

According to Kuran, the tradeoff between the three types of utility allows for a revelation of an agent's preference to be seen as what is revealed by public opinion. It further allows us to represent whether an agent's revealed preference reflects preference falsification, and to what degree her private preference is being falsified. In the context of a social norm, a common individual's gain over her intrinsic utility is independent from her point of view when deciding on what to do when responding to a social norm. This is because opinions of common people don't have substantial social influence over a social outcome reflecting a social norm. Hence, intrinsic utility can be disregarded in analysing behaviour regarding social norms. However, we can see that reputation utility plays the key role in determining what public preferences to reveal because it is to do with the gain from social approval. The heavier normative or moral weight a social norm carries, the higher reputational utility will be at risk in individuals' manifested public opinions governed by that norm. Furthermore, as expressive utility is associated with the publicity of agents' opinions to be expressed, we can see that it is the tradeoff between the reputational utility and expressive utility which matters the most for an agent to determine whether to follow a social norm or not. For example, we can imagine a fully compliant norm follower sacrificing all her expressive utility but gaining the highest possible reputational utility by publicly supporting an option she dislikes, whereas a committed "rebel" would gain highest expressive utility but lose all her reputational utility where behaviour governed by that norm is concerned.

Kuran defines the distribution of public preferences across individuals in a society as “public opinion”, and the distribution of private preferences as “private opinion”. Kuran also argues that the observability of public opinions results from the distribution of observable individual choices, whereas private opinions are less easy to measure. The public opinion constituted by individual public preferences may transform itself through changes of individual choices. Some distributions of private and public preferences may be equilibria in the sense of being relatively stable – no one would improve their net utility by revealing more private preferences that conflict with public ones, or more public opinions that conflict with their private ones.

In studies of social norms for the purpose of norm identification, it is the public’s opinion and how to identify public preferences through observable behaviours of individuals which is of interest. Following Kuran’s analysis, we can say that preference falsification will *always* exist in social dynamics to some degree: the more people seek social approval, the more likely they are to falsify some private preferences in public. The less social approval matters, the less likely individuals are to convey falsified preferences. What matters are the distribution of points along the spectrum of publicity of preference different people find associated with the highest utility.

Cases modelled by Kuran are superficially similar to cases modelled as Pluralistic Ignorance (PI) in the social psychology literature. Miller and McFarland (1987) argue that PI occurs when individuals infer that identical actions of the self and others reflect different internal states. Latana and Darley (1970) use the PI model to analyse the bystander effect that people act similarly to others but assume that their perceptions must be different from those of others. In the typical college drinking example in the literature on PI (Katz & Allport, 1931), the driver for the phenomena is not the fact that the typical college student has falsified ex-ante preferences about how much alcohol she *should* drink at a social gathering event

with other peers. Rather, her behaviour manifests PI if her behaviour is influenced by the descriptive norm she observes.

Bicchieri applies her theory of conditional norms to cases of PI (Bicchieri 2006, p. 186; 2017, p. 47). Bicchieri interprets PI as "a psychological state characterised by the belief that one's private thoughts, attitudes, and feelings are different from those of others, even though one's public behaviour is identical" (2006, p. 186). This analysis seems to conflate preference falsification with PI. Bicchieri (2017, p. 42) demonstrates the concept of PI by an example of a UNICEF study about violence against children. The study shows that a large number of caregivers report a negative judgment on corporal punishment being applied to children, but go on inflicting it on their own children. Bicchieri's explanation of the case relies on people's fear of being regarded as weak or uncaring.

This analysis conflicts with her emphasis on pluralistic ignorance. Instead, this case could be modelled by preference falsification: the reported negative judgments are personal preferences, and the punishing behaviour shows a public preference caregivers reveal in public. That is to say, in cases modeled as preference falsification, what differs from the PI model is that an agent can distinguish between her private normative belief and public opinion.

In cases of preference falsification a person *ex-ante* can make a distinction between her private normative belief and what she thinks is public opinion. In a PI case the problem is precisely that she can't do this because her descriptive beliefs are false. PI arises more easily than preference falsifications and is easier to correct.

Cases featuring preference falsification can *also* feature PI. Take the example of preference falsification on gay marriage. In debates about gay rights and the normative place of gays in society, one of the areas where PI can arise is with respect to people's beliefs about how many people are gay. This will influence individuals' descriptive expectations. Forty years ago in Western society, most people underestimated how many gay people there were. That likely influenced

people's views about gay marriage. If people thought there were hardly any gay people, then invalid inferences to the effect that whatever is statistically strange must be normatively wrong would lead people to think being gay was deviant. By contrast, in the current era, if we ask people from North America what proportion of people are gay, most people will overestimate it (Newport, 2015). What's going on here? One possibility is that as the society isn't normatively against gay marriage anymore, far fewer people view being a gay as deviant. However, there is another more modest hypothesis, if we think about it following the PI model. In the 1960s or 1970s, most gay people were in the closet, so people didn't encounter as many people saying that they were gay as often as they do now. This issue is one of representativeness bias. In this example, PI exists in both times. But there was likely only preference falsification during the period of transition in public opinion.

As Kuran argues, sudden swings of public opinion signify the existence of preference falsification. The historical case study of racial affirmative action in the United States used in Kuran's work in demonstrating his theory is a good example to support this argument. However, PI cases are revealed differently. As PI gets corrected, what society tends to get is gradual incremental reform of behaviour. Take the college drinking case as an example. Gradually more and more college students have learned about statistical drinking norms. This causes students' behaviour to be adjusted gradually to come into line with reality.

As to what matters to the present analysis, a key difference between preference falsification and PI is that in PI cases it is the false descriptive beliefs that need to be addressed, whereas with preference falsification, it is the falsified normative beliefs that are relevant. That's what gives rise to the tradeoff between reputational utility and expressive utility. Therefore, Bicchieri does not address preference falsification when she addresses PI.

A reconciliation of Kuran's and Bicchieri's analyses would require an agreement on the philosophical foundations underlying the accounts, in terms of the

mentalistic concepts to which they appeal. I have argued in Chapter 3 that the alignment between the Intentional Stance and Revealed Preference Theory is the best philosophical framework for interpreting Bicchieri. However, the private/public preference dichotomy in Kuran's model misaligns with this philosophical framework. On the other side, Bicchieri's distinction between moral and social norms is also problematic from this perspective.

First, consider Kuran's concept of private preference.

The concept of private preference is against the Dennettian account of the Intentional Stance (IS), which I have suggested as the philosophical framework for understanding Bicchieri's theory of social norms. According to the IS, propositional attitudes such as beliefs and preferences are abstract posits that help us to track patterns in observable behaviour, including linguistic behaviour. The IS suggests that all there is to having a belief that *p* is being a system that is predictable under the assumption that it believes *p*. That is to say, mentalistic concepts such as beliefs are real patterns which we attribute in order to make sense of others' (and of our own) behaviour, and to predict others' (and our own) actions.

The concept of private preference also misaligns with RPT. According to RPT, preferences are defined in terms of choice behaviour. That is to say, inference to an individual's utility function is based on observed consistent behaviour. If there were true private preferences, then they wouldn't be observable, hence couldn't be modelled by a utility function grounded in RPT.

Yet, we should not deny the existence of *some* kind of private preferences: people have *secret* preferences over some habits which are never revealed to anyone. For example, a habit of wiping oneself from front to back after defecation is a pure habit many people aren't even aware of. No one cares about others' habits of this kind, or even observe them, so this habitual behaviour generally provides zero social influence on anyone else.

Can a concept of secret preference make sense from the perspective of the IS?

Secret habits are truly private. But are they *preferences*, according to the IS? Imagine two cases with respect to the bum-wiping habit. Assume case 1 in which a person who wipes herself from front to back but is not conscious of this behaviour. In this case, it would be misleading to say that she has a preference. This is because she doesn't frame her behaviour in terms of an ascribed preference. But someone else might be aware of their habit and might explicitly frame it to themselves as how it's best to clean themselves, or even, in as many words, as how they prefer to do so. Then this case 2 person applies the IS to herself in this aspect. It is the application of framing using typically sotto voce public language that gives rise to preference in the second case. Application of language is therefore use of social scaffolding. According to the IS the mind itself is an interface pattern that describes the systematic relationship between brains and socially scaffolded minds. Thus a replacement of the concept of private preference by the concept of secret preference could make sense under the IS. However, if secret preference is what Kuran truly means by private preference, then there couldn't be of any importance to the social phenomena he intends to model. When Kuran talks about private preferences, he must not mean secret preferences. Secret preferences are not patterns of socially integrated behaviour, and they are of no social significance. Social significance is the key characteristic in the modelling of preference falsification following the utility model developed by Kuran. Genuinely secret preferences carry no social significance.

So, what kind of non-secret preference that matters to Kuran's account would be counted as private? Kuran says that privacy can be understood as varying in degrees along a spectrum. A preference may appear to be private for some individuals when interacting with some people, but not when interacting with others. What determines the degree of privacy is the targeted social groups that matter to the individuals in question. For example, a relatively strongly private preference might be a preference that is revealed in interacting with a person's partner, or immediate

family member. A less private preference might be one that emerges in interactions in one's pub but would not be revealed in a letter to the editor.

The distinction between personal and social belief in excluding moral norms from the scope of social norms in Bicchieri's analysis likewise is in tension with the Intentional Stance. As reviewed in Chapter 2, a key ingredient in Bicchieri's theory is the concept of belief and expectation. The IS suggests that people form belief/expectation in order to coordinate with others in society. Moralised beliefs and preferences are what lead to the most important expectations we have in social interactions, so should also be interpreted as coordination devices. Propositional attitude ascriptions, whether moralised or not, pick out relations among an agent, features of her environment, and patterns of her social expectations.

This view gains support from research in social science that moral beliefs arise from the same processes of socialisation (i.e., learning, imitation, pedagogy) as other norms. For example, according to Binmore (1994, 1998, 2005), "something is morally right when an approved strategy has been followed in sustaining the current equilibrium. Something is morally good when an equilibrium has been selected in the manner approved by the current social contract" (2005, p. 96). Binmore (2010) rejects the distinction between moral and social norms and argues that 'social norm' is a general term which includes moral norms, legal norms and any other social facts about instituted rules.

In the tradition of philosophical debates about morality, it is implausible that a completely personal norm could be reasonably moralised apart from the attitudes of any others in society. A person who announced a completely idiosyncratic moral principle would be regarded as not understanding the concept of morality.

There are three kinds of broad views about the concept of moral beliefs. The first broad view sees moral beliefs as innate in human nature. This has been a common theme since the ancient wisdom in Confucianism, and has had some defenders in Western philosophy. With the rise of modern empirical science, it still

remains popular. Philosophical arguments and statistical data have provided positive evidence about the innateness of human pro-sociality (Hamilton 1964; Henrich & Boyd 2001). However, this does not imply innate specific morality. Joyce (2006) reviews empirical evidence and concludes that "mechanisms of cultural transmission play an enormous and perhaps exhaustive role in determining the content of an individual's moral convictions" (p. 140).

A second kind of view which is popular among Western philosophers is the Categorical Imperative (CI) developed by Immanuel Kant (1724-1804) who argued that the principle of morality can be discovered by practical rationality. Kant argues that the CI is an objective, rationally necessary and unconditional principle that should be followed by everyone despite any natural desires that humans have to the contrary. On the one hand, Kant argues that everyone can acquire this principle through the practice of practical reasoning; on the other hand, conformity to the CI is essentially necessary for rational agency. Suppose Kant were right, and you were the only rational agent? Then you'd do the Kantian reasoning all by yourself to gain moral beliefs. Also, according to the CI, rational agents' acts and beliefs are only counted as moral if they result from practical reasoning, instead of social influences. Thus for Kant moral beliefs are outcomes of rational reflections, which do not root in society, but in individual human autonomy (Kant 1785).

Lawrence Kohlberg's cognitive-development theory which followed Jean Piaget's work on children's moral development, was Kantian in holding that human moral reasoning in principle involves six developmental stages, but that most people fall short of the "fully moral" sixth stage. Kohlberg's work was principally concerned with people's cognitive development of an understanding of justice. More recent empirical work in evolutionary psychology entirely rejects such Kantian moral psychology.

A third view is that morality is naturally socially constructed. For example, the social philosopher Mead (1925) takes normative and moral evaluation to be an

indispensable facet of social existence, and argues that the emergence of a moral self is a product of interactional engagement within complex social contexts. Mead argued that the process of emerging as a self embeds individuals within the normative expectations and values of others in their social surroundings that "the structure of self expresses or reflects the general behavior pattern of this social group to which he belongs" (1934, p. 7). Moral psychologist Johnathan Haidt argues that virtues are social skills that "to possess a virtue is to have disciplined one's faculties so they are fully and properly responsive to one's local sociomoral context" (2007, p. 61). Haidt & Joseph (2004) and Haidt (2012) report cross-cultural empirical evidence for the social shaping of moral beliefs, an interpretation that is supported by empirical findings from Nisbett (2004) and Flanagan (2017).

The key point here is that if morality were arrived by pure reason or were innate, then people could have private moral commitments. But these views have been empirically refuted. As we saw earlier in this chapter, Bicchieri (2014) accepts this conclusion from moral psychology.

Kuran cases then provide further reasons to reject the exclusion of moral norms from the sphere of social norms as suggested by Bicchieri. On Kuran's model, the heavier normative weight a social norm carries, the higher the likelihood of preference falsification and the heaviest possible normative weight is precisely what moralisation of a norm signals.

The discussion above uncovers a problem in Bicchieri's theory of conditional norms. On the one hand, Bicchieri's analysis of conditional norms relies on the distinction that moral norms are unconditional. On the other hand, it seems that in order to account for Kuran cases and preference falsification and for empirical applicability, her theory must drop this distinction and incorporate moral norms into her analysis by assuming moral norms are also conditional.

Based on the discussion above, I suggest a modification in both Kuran's and Bicchieri's theories in order to reconcile them. Bicchieri's account should be

amended to recognise moral norms as also conditional, as also affecting individuals' normative behaviour via joint effects of expectations and preferences. Kuran's private/public preference dichotomy should be reconceptualised as between minority and majority preferences. Minority preferences are preferences applied for coordinating with people in one, relatively intimate, reference group for an agent, whereas majority preferences are preferences applied for coordination in a different and larger reference group, where relative to that agent's social position, public opinion is identified.

The modification of both theories on the basis of the philosophical framework suggested in Chapter 3 leads to a basis for reconciliating them. It requires an application of the concept of reference network discussed in Bicchieri's work. In Bicchieri's analysis, a reference network refers to the set of people whose actions and beliefs matter for shaping agents' social expectations.

It seems that once we apply the concept of a reference network, we can confirm the following. First, it seems that Bicchieri's theory of conditional norms does not have to exclude the phenomenon of preference falsification by sticking to an empirically implausible distinction between moral norms and conditional social norms. Second, the misleading concept of private preference in Kuran's model can be replaced by the concept of misalignment of reference networks, i.e., minority versus majority preference. Third, the application of the idea of a reference network here is consistent with Bicchieri's view that the existence of belief consistency at the individual level is one of the conditions for norm identification within one reference network. Fourth, it allows Bicchieri's theory of conditional norms and Kuran's model of preference falsification to be reconciled for analysing norm compliant behaviours.

By emphasising the concept of reference network, we therefore can see that Bicchieri's model should be understood as an analysis of norms in general in a community where everyone shares the same reference network. This coincides with

Bicchieri's own understanding of social norms, according to which "the conditions for a norm to exist entail, when they are fulfilled, that a social norm is an equilibrium. It is a situation of stable mutual adjustment: Everyone anticipates everyone else's behaviour, and all these anticipations turn out to be correct. Social norms have no reality other than our beliefs that others behave according to them and expect us to behave according to them. In equilibrium, such beliefs are confirmed by experience and thus they become more and more ingrained as time goes on." (Bicchieri 2006, p. 22-23).

The importance of this finding is not only conceptual. The practical implications are as follows. When adopting Bicchieri's model for empirical modelling of social norms, the first step must be experimental control for reference network alignment. If there exists a social norm of relevant expectations/beliefs that align, the simplest case for application of Bicchieri's theory arises where we expect the norms of interest (i.e., behaviour in an Ultimatum Game, or Trust Game), in one reference network shared by everyone. In fact, in her discussion of measuring and changing norms, Bicchieri emphasises the same procedure that "mapping the reference network is an essential part of understanding social norms and how to change them, because the norm has to change within the reference network (2017, p. 53)." In cases where there are multiple reference networks at play, we must expect consequent complexity of normative behaviour. In particular, we should expect to see preference falsification.

7.2 Further Extensions for Future Work

Another extension of the work in this thesis would involve moving beyond the evaluation of the beliefs of the complete group of subjects, as if they were homogenous, to allowing for individual beliefs. The assumption of homogeneity was appropriate to test the general concept of social norms proposed by Bichierri, but that concept also allows for the relevant reference group to be a sub-group of the

population. The experimental design already provides the individual beliefs of each subject, since it employed a procedure to "risk neutralise" responses to the scoring rule used. Therefore, theory tells us that the observed reports for each subject *are* their beliefs. One could technically ask if there exist arbitrary sets of individuals of a given size for which the hypothesis of a social norm is accepted, but that is not a particularly interesting question. Instead, an interesting question would be to evaluate stratified sub-groups in terms of observed demographic characteristics (e.g., men or women, or racial groups) or responses to the values survey (e.g., attitudes towards the role of government), and see if those sub-groups have beliefs that are consistent with a social norm.

A related extension would be to use Bayesian methods to formulate the hypotheses of a social norm in terms of belief consistency. The same model used in Chapter 6, the heteroskedastic ordered probit model, can be directly formulated as a Bayesian model.[37] The great advantage of this extra statistical step is that the core hypotheses can then be defined in terms of Regions of Practical Equivalence, usefully called a ROPE (Kruschke 2015, 2018). A ROPE is just an interval hypothesis around the point hypotheses defining a social norm in terms of two belief distributions having the same mean and variance. By varying the tightness of the ROPE around the point hypotheses, one can ascertain the posterior probability of the hypothesis of a social norm for various levels of the ROPE. Intuitively, and formally, if the ROPE is loose enough, then *any* observed behavior can be consistent with a social norm, since it is consistent with "anything". But as one tightens the ROPE, at what levels can one accept the hypothesis of a social norm? In this case Bayesians often use the 50% posterior probability level, since it refers to a preponderance of the evidence. But usually one displays the mapping from various ROPE tightness levels to implied posterior probability of the hypothesis being accepted, allowing the

[37] In *Stata* the classical estimation command **hetoprobit** is just re-estimated using the **bayes: hetoprobit** command.

reader to make her own mind up about the evidence.

A final extension would connect the last two extensions: use the statistical methods to find certain sub-groups in terms of identifiable characteristics that are *most* likely or *least* likely to have social norms. Then bring those sub-groups back into the lab, perhaps even just sub-groups from the same population, and allow them to play the UG. One would expect to see different outcomes in these UGs if the hypothesis about social norms is valid. A related extension would be to provide some information to subjects about the beliefs of sub-groups and allow them to "self-select" into groups that would then play the UG (as pairs). The idea of self-selection is well known in economics from the concept of a Tiebout equilibrium, where individuals or households can "vote with their feet" and migrate to locations that offer more of the characteristics, such as taxes and public goods, that they prefer (Botelho et al. 2022). Normative reference networks often form under such dynamics.

References

Andersen, S., Fountain, J., Harrison, G. W., & Rutström, E. (2014). Estimating subjective probabilities. *Journal of Risk and Uncertainty* 48(3): 207-229.

Ashraf, S., Kuang, J., Das, U., & Bicchieri, C. (2020). Sanitation practices during early phases of COVID-19 lockdown in Peri-urban communities in Tamil Nadu, India. *American Journal of Tropical Medicine and Hygiene* 103(5): 2012-2018.

Ayoya, M. A., Bendech, M. A., Zagre, N. M., & Tchibindat, F. (2012). Maternal anaemia in West and Centra Africa: time for urgent action. *Public Health Nutrition* 15(5): 916-927.

Benoît, J. P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica* 79(5): 1591-1625.

Bicchieri, C. (2006). *The Grammar of the Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

---. (2010). Norms, preferences, and conditional behavior. *Politics, Philosophy & Economics* 9(3): 297-313.

---. (2014). Norms, conventions, and the power of expectations. In *Philosophy of Social Science: A New Introduction.* Edited by Cartwright, N. & E. Montuschi. Oxford: Oxford University Press., pp. 208-230.

---. (2017). *Norms In The Wild: How To Diagnose, Measure, And Change Social Norms*. New York: Oxford University Press.

---. (2022). Norm nudging: how to measure what we want to implement. In *Behavioral Science in the Wild*. Edited by Mazar, N. & D, Soman. Toronto: University Toronto Press.

Bicchieri, C., & Dimant, E. (2019). Nudging with care: the risks and benefits of social information. *CeDEx Discussion Paper Series 2019-02,* Center for Decision Research and Experimental Economics (CeDEx), University of Nottingham.

Bicchieri, C., & Dimant, E., Sonderegger, S. (2019). It's not a lie if you believe it: On norms, lying, and self-serving belief distortion. *CeDEx Discussion Paper Series 2019-07*, Center for Decision Research and Experimental Economics

(CeDEx), University of Nottingham.

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: public information and fairness norms. *Journal of Behavioral Decision Making* 23: 161-178.

---. (2013). Norm manipulation, norm evasion: experimental evidence. *Economics & Philosophy* 29(2): 175-198.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making* 22: 191-208.

--- (2010). When equality trumps reciprocity. *Journal of Economic Psychology* 31: 456-470.

Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy & Economics* 10(2): 170-187.

Bicchieri, C., & Zhang, J. (2012). An Embarrassment of riches: modeling social preferences in ultimatum games. In *Philosophy of Economics*: 577-595. Edited by Mäki, U. San Diego: North Holland.

Binmore, K. (1994). *Game Theory And The Social Contract, Volume 1: Playing Fair.* Cambridge MA: MIT Press

---. (1998). *Game Theory And The Social Contract, Volume 2: Just Playing.* Cambridge MA: MIT Press

---. (2005). *Natural Justice*. Oxford: Oxford University Press.

---. (2006). Why do people cooperate? *Politics, Philosophy & Economics* 5(1): 81-96.

---. (2007). *Game Theory: A Very Short Introduction*. Oxford: Oxford University Press.

---. (2009). *Rational Decisions*. Princeton: Princeton University Press.

---. (2010). Social norms or social preferences? *Mind & Society* 9: 139-157.

Blackburn, M., Harrison, G. W., & Rutström, E. E. (1994). Statistical bias functions and informative hypothetical surveys. *American Journal of Agricultural Economics* 76(5): 1084-1088.

Blumenschein, K., Johannesson, M., & Yakoyama, K. (2001). Hypothetical vs. real willingness to pay in the health sector: results from a field experiment. *Journal of Health Economics* 20(3): 441-457.

Botelho, A., Harrison, G. W., Pinto, L., Ross, D., & Rutström, E. (2022). Endogenous choice of institutional punishment mechanisms to promote social cooperation. *Public Choice* 191: 309-335.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 75: 1-3.

Burge, T. (1986). Individualism and psychology. *The Philosophical Review* 95(1): 3-45.

Burks, S. V., & Krupka, E. L. (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: evidence from the financial services industry. *Management Science* 58(1): 203-217.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree – an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9: 88-97.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge MA: MIT Press.

---. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Crawford, S. & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review* 89(3): 582-601.

Cummings, R. G., Elliott, S., Harrison, G. W., & Murphy, J. (1997). Are hypothetical referenda incentive compatible? *Journal of Political Economy* 105(3): 609-621.

Cummings, R. G., & Taylor, L. O. (1998). Does realism matter in contingent valuation surveys? *Land Economics* 74(2): 203-215.

Dana, J., Cain, D., & Dawes, R. (2006). What you don't know won't hurt me: costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes* 100(2): 193-201.

Dana, J., Weber, R., & Kuang, X. (2007). Exploiting moral wiggle room:

experiments demonstrating an illusory preference for fairness. *Economic Theory* 33(1): 67-80.

Dennett, D. (1987). *The Intentional Stance*. Cambridge MA: MIT Press.

---. (1988). Out of the armchair and into the field. *Poetics Today* 9(1): 205-221.

---. (1991). Real patterns. *The Journal of Philosophy* 88(1): 27-51.

Durkheim, E. (1982). *The Rules of Sociological Method*, tr. W. D. Halls. New York: Free Press.

Engelmann, D. & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review* 94(4): 857-869.

Epstein, B. (2015). *The Ant Trap.* Oxford: Oxford University Press.

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009-2015. *Educational Researchers* 47: 221-234.

Fehr, E., & Gächter, S. (2000a). Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives* 14(3): 159-181.

---. (2000b). Cooperation and punishment in public good experiments. *American Economic Review* 90(4): 980-994.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature* 422(6928): 137-140.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114(3): 817-868.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour* 2(7): 458-468.

Flanagan, O. (2017). *The Geography of Morals*. New York, Oxford University Press.

Fodor, J. (1996). Deconstructing Dennett's Darwin. *Mind & Language* 11(3): 246-262.

Gächter, S., Nosenzo, D. & Sefton, M. (2013). Peer effects in pro-social behavior:

social norms or social preferences? *Journal of the European Economic Association* 11(3): 548-573.

Gilbert, M. (1989). *On Social Facts*. Princeton: Princeton University Press.

Gintis, H. (2009). *The Bounds of Reason*. Princeton: Princeton University Press.

Goyal, S. (2007). *Connections: An Introduction to the Economics of Networks*. Princeton: Princeton University Press.

Guala, F. (2016). *Understanding Institutions*. Princeton: Princeton University Press.

---. (2022). Rescuing ontological individualism. *Philosophy of Science* 89: 471-485.

Haidt, J. (2007). The New Synthesis in Moral Psychology. Science 316: 998-1002.

---. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* New York, Vintage Books.

Haidt, J. & Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(3): 55-66.

Hamilton, W. D. (1964). The Genetical Evolution of Social Behaviour. *Journal of Theoretical Biology* 7: 1-16.

Hands, W. (2013). Foundations of contemporary revealed preference theory. *Erkenntnis* 78(5): 1081-1108.

Harcourt-Cooke, C., Els, G., & Van Rensburg, E. (2022). Using comics to improve financial behaviour. *Journal of Behavioral and Experimental Finance* 33: 100614.

Harrison, G. W. (2006a). Making choice studies incentive compatible. In *Valuing Environmental Amenities Using Stated Choice Studies*. Edited by Kanninen, B. J. Dordrecht: Springer.

---. (2006b). Experimental evidence on alternative environmental valuation methods. *Environmental & Resource Economics* 34: 125-162.

---. (2014). Real choices and hypothetical choices. In *Handbook of Choice Modelling*. Edited by Hess, S., & Daly, A. Cheltenham: Edward Elgar.

Harrison, G.W., Hofmeyr, A., Kincaid, H., Monroe, B., Ross, D., Schneider, D., &

Swarthout, T. J. (2021). Eliciting beliefs about COVID-19 prevalence and mortality: epidemiological models compared with the street. *Methods* 195: 103-112.

Harrison, G. W., Martínez-Correa, J., & Swarthout, T. J. (2013). Inducing risk neutral preferences with binary lotteries: a reconsideration. *Journal of Economic Behavior & Organization* 94: 145-159.

---. (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization* 101: 128-140.

Harrison, G. W., Martínez-Correa, J., Swarthout, T. J., & Ulm, E. R. (2015). Eliciting subjective probability distributions with binary lotteries. *Economics Letters* 127: 68-71.

---. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization* 134: 430-448.

Harrison, G. W., & Ross, D. (2016). The psychology of human risk preferences and vulnerability to scare-mongers: experimental economic tools for hypothesis formulation and testing. *Journal of Cognition and Culture* 16: 383-414.

Harrison, G. W., & Swarthout, T. (2021). Belief Dsitributions, Baye's Rule and Bayesian Overconfidence. *Working Paper 2020-11*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Hausman, D. M. (1992). *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

---. (2000). Revealed preference, belief, and game theory. *Economics & Philosophy* 16(1): 99-115.

---. (2008). Mindless or mindful economics: a methodological evaluation. In *Foundations of Positive and Normative Economics: A Handbook*. Edited by Caplin, A., & Schotter, A. Oxford: Oxford University Press.

Henrich, J. & Boyd, R. (2001). Why People Punish Defectors Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. *Journal of Theoretical Biology* 208: 79-89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., et al. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28(6):

795-815.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., et al. (2006). Costly punishment across human societies. *Science* 312(5781): 1767-1770.

Hossain, T., & Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies* 80: 984-1001.

Joyce, R. (2006). *The Evolution of Morality.* London: MIT Press.

Kadane, J. B., & Winkler, R. L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association* 83(402): 357-363.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*, tr. H. Paton. New York: Harper & Row.

Katz, D. & Allport, F. (1931). Student Attitudes. *Syracuse*: 1-8. NY: Craftsman Press.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association* 14(3): 608-638.

---. (2018). A portable method of eliciting respect for social norms. *Economic Letters* 168: 147-150.

---. (2020). A Theory of injunctive norms. *Mimeo*, Chapman University and Maastricht University. http://dx.doi.org/10.2139/ssrn.3566589

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of the Economic Association* 11(3): 495-524.

Kruschke, J. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. New York: Elsevier (Second Edition).

---. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science* 1: 270-280.

Kuran, T. (1990). Private and Public Preferences. *Economics and Philosophy*. 6(1): 1-26.

---. (1995). *Private Truths, Public Lies*. Cambridge: Harvard University Press.

Ladyman, J., & Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Latana, B., & Darley, J. (1970). *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Croft.

Lazear, E., Malmendier, U., & Weber, R. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics* 4(1): 136-163.

List, J. A. (2001). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sports cards. *American Economic Review* 91(5): 1498-1507.

---. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy* 115(3): 482-493.

Martin, J. L. (2009). *Social Structures*. Princeton: Princeton University Press.

Matheson, J. E., & Winkler. R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* 22(10): 1087-1096.

Mead, G. H. (1925). The Genesis of the self and social control. *International Journal of Ethics* 35: 251-277.

---. (1934). *Mind, Self, and Society* (Vol. 111). Chicago: University of Chicago Press.

Merkle, C. & Weber, M. (2011). True overconfidence: the inability of rational information processing to account for apparent overconfidence. *Organizational Behavior and Human Decision Processes* 116(2): 262-271.

Miller, D. & McFarland, C. (1987). Pluralistic Ignorance: When Similarity is Interpreted as Dissimilarity. *Journal of Personality and Social Psychology* 53 (2): 298-305.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review* 115(2): 502.

Muldoon, R. (2022). Social norms. In *The Routledge Handbook of Philosophy, Politics, and Economics*. Edited by Melenovsky, C. M. New York: Routledge.

Newport, F. (2015). Americans greatly overestimate percent gay, lesbian in US. *Gallup News, May, 21*.

Nisbett, R. (2004). *The geography of thought: How Asians and Westerners think differently…and Why*. New York: The Free Press.

North, D. C. (1991). Institutions. *Journal of Economic Perspectives* 5(1): 97-112.

---. (1993). The new institutional economics and development. *Economic History 9309002*: 1-8.

Ostrom, V. (1980). Artisanship and artifact. *Public Administration Review* 40: 309-317.

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics* 42: 3-45.

Ross, D. (2000). Rainforest realism: a Dennettian theory of existence. In *Dennett's Philosophy: A Comprehensive Assessment:* 147-168. Edited by Ross, D., Brook, A., & Thompson, D. Cambridge MA: MIT Press.

---. (2005). *Economic Theory and Cognitive Science: Microexplanation*. Cambridge MA: MIT Press.

---. (2006). Moral Functionalism, Preference Moralization, and Anti-Conservatism: Why Metaethical Error Theory Doesn't Imply Policy Quietism. *Social Network Research Network*: 1-30. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=962447

---. (2011). Naturalism: the place of society in nature. In *The Sage Handbook of Philosophy of Social Science:* 147-168. Edited by Jarvie, I., & Zamorra-Bonilla, J. London: Sage Publications.

---. (2014). *Philosophy of Economics*. Houndmills, Basingstoke: Palgrave Macmillan.

---. (2022). Economics is converging with sociology but not with psychology. *Journal of Economic Methodology*: 1-22. https://doi.org/10.1080/1350178X.2022.2049854

Ross, D., Stirling, W. C., & Tummolini, L. (2021). Modeling norm-governed communities with conditional games. *CEAR Working Paper WP 2021-05*, The Center for the Economic Analysis of Risk, Georgia State University.

https://cear.gsu.edu/working-papers/

Roth, A. E., & Malouf, M. W. (1979). Game-theoretic models and the role of information in bargaining. *Psychological Review* 86(6): 574.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica* 15(60): 243-253.

Savage, L. J. (1972). *The Foundations of Statistics.* New York: Dover Publications, Second Edition.

Savellos, E., & Yalçin, Ü., editors (1995). *Supervenience: New Essays.* Cambridge, U.K.: Cambridge University Press.

Schmidt, R. J. (2019). Do injunctive or descriptive social norms elicited using coordination games better explain social preferences? *Discussion Paper Series, 668,* University of Heidelberg. https://doi.org/10.11588/heidok.00027175

Schram, A., & Charness, G. (2015). Inducing social norms in laboratory allocation choices. *Management Science* 61(7): 1531-1546.

Sen, A. (1973). Behaviour and the concept of preference. *Economica* 40(159): 241-259.

---. (1980). Description as choice. *Oxford Economic Papers* 32(3): 353-369.

---. (1993). Internal consistency of choice. *Econometrica* 61(3): 495-521.

---. (1997). Maximization and the act of choice. *Econometrica* 65(4): 745-779.

Simmel, G. (1895). The Problem of sociology. *The Annals of the American Academy of Political and Social Science* 6(3): 52-63.

Smith, C. A. B. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society: Series B (Methodological)* 23(1): 1-25.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review* 72(5): 923-955.

---. (2007). *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge UK: Cambridge University Press.

Smith, V. L. & Walker, J. M. (1993). Monetary rewards and decision cost in

experimental economics. *Economic Inquiry* 31(2): 245-261.

Sneddon, A. (2011). *Like-Minded: Externalism and Moral Psychology*. Cambridge MA: MIT Press.

StataCorp. (2021). *Stata: Release 17. Statistical Software*. College Station, Texas: StataCorp

Sugden, R. (1998). Normative expectations: the simultaneous evolution of institutions and norms. *Economics, Values, and Organization* 73.

---. (2004). *The Economics of rights, Co-operation, and Welfare*. Basingstoke: Palgrave Macmillan.

Tastle, W. J., & Wierman, M. J. (2007). Consensus and dissention: a measure of ordinal dispersion. *International Journal of Approximate Reasoning* 45(3): 531-545.

Tversky, A. (1997). Features of similarity. *Psychological Review* 84(4): 327-352.

Tversky, A. & Gati, I. (1978). Studies of similarity. Cognition and Categorization 1(1978): 79-98.

Von-Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, Second Edition.

Wolf, M. & Koons, J. R. (2016). *The Normative and the Natural*. Houndmills, Basingstoke: Palgrave Macmillan.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association* 67(337): 187-191.

Winkler, R. L. & Murphy, A. H. (1970). Nonlinear utility and the probability score. *Journal of Applied Meteorology and Climatology* 9(1): 143-148.

Xiao, E., & Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology* 31(3): 456-470.

Xiao, E., & Houser, D. (2009). Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange. *Journal of Economic Psychology* 30(3): 393-404.

Zahle, J., & Kincaid, H. (2019). Why be a methodological individualist? *Synthese*

196: 655-675.

Zawidzki, T. (2007). *Dennett*. Oxford: Oneworld.

---. (2012). Unlikely allies: embodied social cognition and the intentional stance. *Phenomenology and the Cognitive Sciences* 11(4): 487-506.

---. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition.* Cambridge MA: MIT Press.

Appendix A: Ultimatum Game Task Instruction

In this task you will be asked to make a decision about dividing an amount of money between yourself and another person. You may also receive money as a result of the decision of another person. In this task there are two roles, referred to as **Proposer** and **Responder**. You and the other participants in this study will be asked to make decisions both as Proposer and as Responder.

After you and all the other participants have made decisions in both roles, you will be randomly and anonymously matched with one of the other participants in the study. All the people participating in this study are UCT students.

After the match is made, a random draw will determine if you will be the Proposer and the other participant the Responder, or the other way around. If you are selected as Proposer, the decision you made in the role of Proposer will be used, while the decisions the other participant made in the role of Responder will be used. The decisions you made in both roles before you were matched, and the random draw that determines which role you will actually take, will determine your earnings.

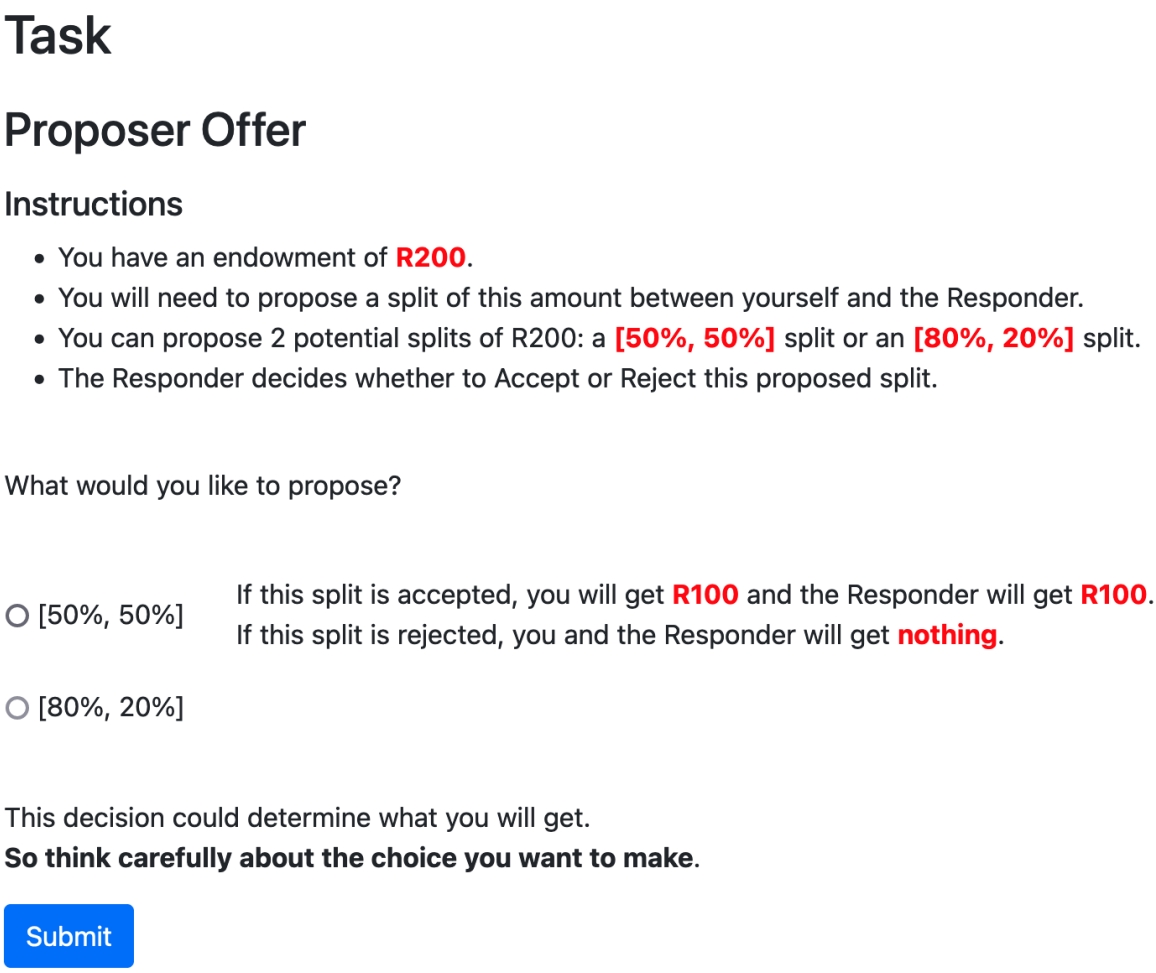
We now describe how the task works. At the start, in the role of Proposer you will be given a monetary endowment, for example R200. You will then be asked to propose a split of this amount between yourself and the Responder. In the role of the Responder, you will be asked to decide, for each potential split of the money that the Proposer can make, whether you will accept or reject the split. If you **accept** the split, you will earn the amount the Proposer offered to you, and the Proposer will earn whatever money they proposed to keep for themself. If you **reject** the split, the Proposer loses the initial endowment of R200 and both of you earn nothing.

The Proposer can only propose two potential splits of the endowment:

1. The Proposer can offer 50% of the endowment to the Responder, while keeping the remaining 50% of the endowment. This implies that if the Proposer has a R200 endowment, the Responder is offered R100 and the Proposer keeps R100.
2. The Proposer can offer 20% of the endowment to the Responder, while keeping the remaining 80% of the endowment. This implies that if the Proposer has a R200 endowment, then the Responder is offered R40 and the Proposer keeps R160.

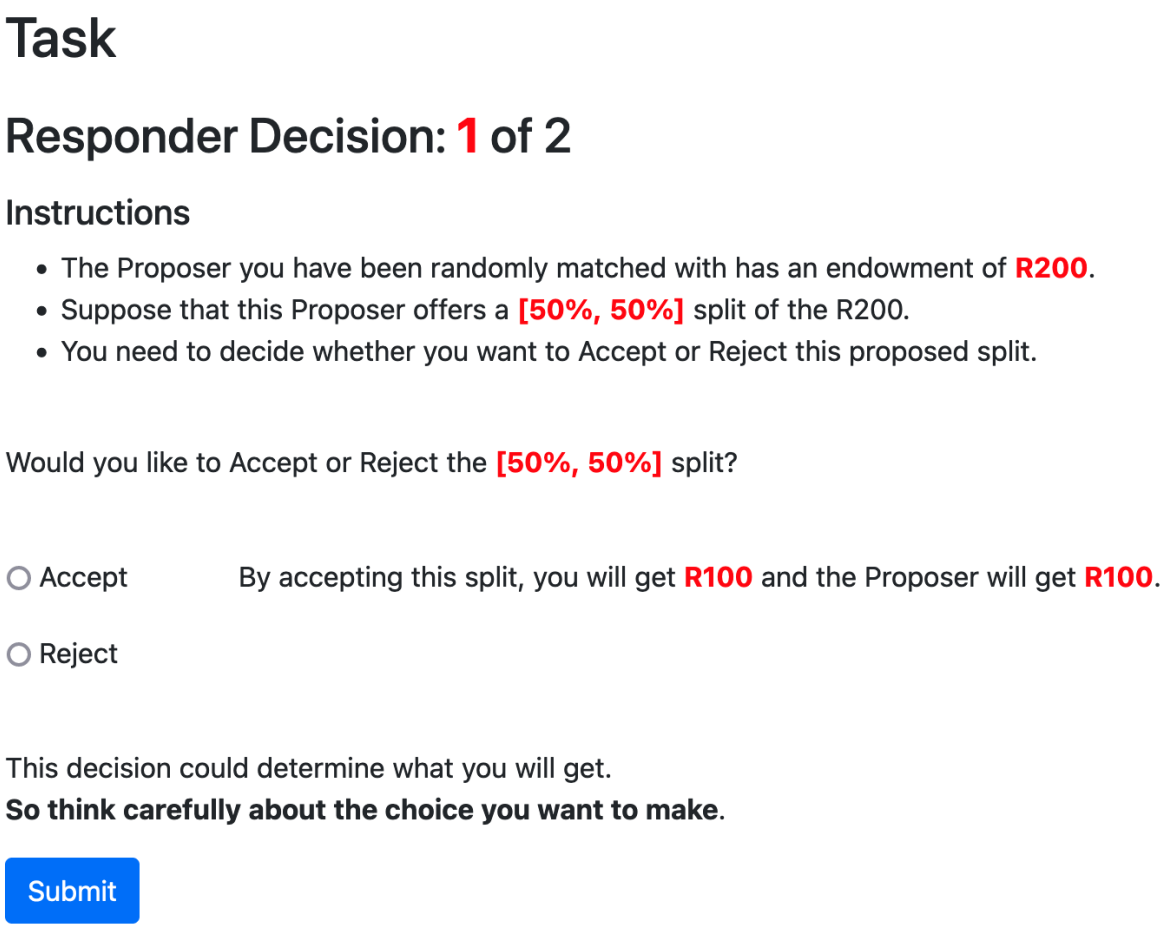
The Responder needs to decide whether to accept or reject the 50% offer if that is the offer they receive, and whether to accept or reject the 20% offer if that is the offer they receive.

This is what the computer display for the Proposer looks like:



As you can see, the display prompts the Proposer to choose between two offers. The first offer is a [50%, 50%] split of R200, corresponding to the Proposer keeping R100, which is 50% of the endowment, and the Responder receiving R100, which is 50% of the endowment. The second offer is an [80%, 20%] split of R200, corresponding to the Proposer keeping R160, which is 80% of the endowment, and the Responder receiving R40, which is 20% of the endowment. The Proposer must decide what split to offer to the Responder, knowing that the Responder will have made a decision whether to accept or reject the [50%, 50%] offer if it is made, and whether to accept or reject the [80%, 20%] offer if it is made. When the Proposer hovers over a choice of [50%, 50%] or [80%, 20%], the Proposer is shown the amounts that the Proposer and the Responder would earn from this choice, as you can see in the screenshot.

This is what the computer display for the Responder looks like:



As you can see, the display shows that the Responder must decide whether to accept or reject an offer of a [50%, 50%] split. If the Responder accepts the [50%, 50%] split, then the Responder receives R100, which is 50% of the endowment, and the Proposer keeps R100, which is 50% of the endowment. If the Responder rejects the offer, then the Proposer loses the R200 endowment, and both the Proposer and Responder earn nothing. When the Responder hovers over a choice of Accept or Reject, the Responder is shown the amounts that the Proposer and the Responder would earn from this choice, as you can see in the screenshot. The Responder must also decide whether to accept or reject an offer of an [80%, 20%] split. If the Responder accepts the [80%, 20%] split, then the Responder receives R40, which is 20% of the endowment, and the Proposer keeps R160, which is 80% of the endowment. If the Responder rejects the offer, then the Proposer loses the R200 endowment, and both the Proposer and Responder earn nothing.

Once you have made your decisions in both the roles of Proposer and Responder the task is over, and your role as either Proposer or Responder will be randomly determined for payment. At the end of the study, we will determine your earnings for this task in the following way:

- You will be randomly and anonymously matched with another participant in the study. You will not know who this person is and they will not know who you are.

- If you have been assigned the role of Proposer, the person you are matched with will have been assigned the role of Responder. On the other hand, if you have been assigned the role of Responder, the person you are matched with will have been assigned the role of Proposer.
- If you are randomly assigned to the Proposer role, the split you offered in that role will be compared to the Responder's decision to accept or reject that specific split. If the Responder chose to accept that split, each of you will be paid the corresponding amounts. If the Responder chose to reject that split, you lose the R200 endowment and both of you earn nothing.
- If you are randomly assigned to the Responder role, your decisions of whether to accept or reject the offer of a [50%, 50%] split and an [80%, 20%] split will be compared to the split that the Proposer offered. If you, as the Responder, chose to accept this split, each of you will be paid the corresponding amounts. If you chose to reject that split, the Proposer loses the R200 endowment and each of you earn nothing.

For example, suppose that you are randomly selected as Proposer, and you chose to offer the [50%, 50%] split, meaning you keep R100 and the Responder receives R100. Suppose that the person you are anonymously matched with chose to accept this split. Then you earn R100 as Proposer, and the Responder also earns R100. If instead the Responder chose to reject this split, you lose the R200 endowment and both of you earn nothing.

As another example, suppose that you are randomly selected as Responder and that the person you are anonymously matched with in the Proposer role chose to offer the [80%, 20%] split, meaning you receive R40 and the Proposer keeps R160. Assume that as the Responder you chose to accept this split. Then, as Responder, you earn R40, and the Proposer earns R160. If you chose to reject this split, then the Proposer loses the R200 endowment and both of you earn nothing.

In this task there are no right or wrong answers. Please make your choices by thinking carefully about the different options you prefer in both the role of Proposer and Responder.

Please click the Next button

Appendix B: Belief Task Instruction for R100 Endowment with Cartoon Treatment

This is a task where you will be paid according to how accurate your beliefs are about the outcomes of an interaction between people, and how accurate your predictions are of other people's beliefs about these outcomes. Your earnings will depend on what the outcomes of the interaction between other people actually are, and on what other people report that they believe about the interaction. You will be presented with some questions and asked to place bets on your beliefs about the answer to each question. You will be rewarded for your answer to one of these questions, so you should think carefully about your answer to each question. The question that is chosen for payment will be determined after you have made all decisions, and that process is described at the end of these instructions. Everyone participating in this study is a UCT student.

The interaction between people, on which you will be asked to place bets, works as follows:

- There are two roles in the interaction: **Proposer** and **Responder**
- The Proposer is given a money endowment of R100, and is asked to propose a split of this amount between themself and the other person, the Responder.
- The split the Proposer can offer is either [50%, 50%] or [80%, 20%] of the R100 endowment, where the first percentage in each potential split is the percentage of the endowment the Proposer would get, and the second percentage in each potential split is the percentage of the endowment the Responder would get. Therefore, with a [50%, 50%] split, the Proposer gets R50 and the Responder gets R50. With an [80%, 20%] split the Proposer gets R80 and the Responder gets R20.
- The Responder will then be asked to decide, for each potential split of the money that the Proposer might offer, whether to accept or reject this proposed split.
- If the Responder **accepts** a proposed split, the Responder receives the amount they were offered, and the Proposer keeps the rest of the money.
- If the Responder **rejects** a proposed split, the R100 endowment is withdrawn, and both the Proposer and the Responder get nothing.

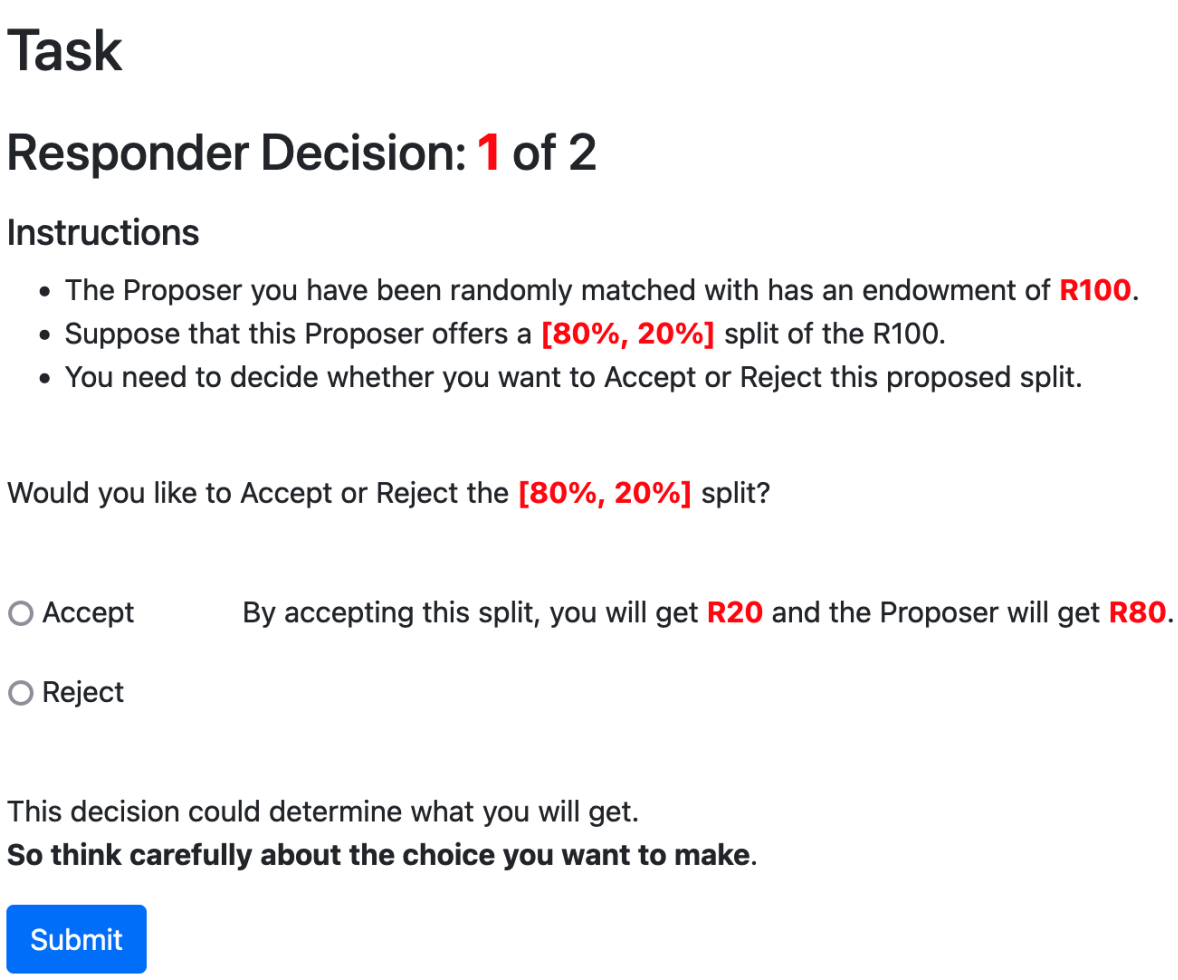
Thus, the Proposer has to make one choice: to propose either a [50%, 50%] split of R100, or an [80%, 20%] split of R100.

By contrast, the Responder has to make two choices: to accept or reject each potential split. If the Responder accepts a proposed [50%, 50%] split of R100, the Proposer gets R50 and the Responder gets R50. If the Responder rejects this split, both the Proposer and the Responder earn nothing from that interaction. If the Responder accepts a proposed [80%, 20%] split of R100, the Proposer gets R80 and the Responder gets R20. If the Responder rejects this split, both the Proposer and the

Responder earn nothing from that interaction.

124 UCT students took part in this interaction recently. Each UCT student assumed the role of Proposer *and* Responder. Once each student had made decisions in both roles, two people were randomly matched, with one person randomly assigned to the role of Proposer, and the other person randomly assigned to the role of Responder. The choices that the Proposer and Responder had made previously were applied to the interaction, and this determined their earnings for the task.

This screenshot shows you the display for the Responder, for the case where the Proposer offers a proposed [80%, 20%] split of R100. If this proposal is accepted, the Proposer gets 80% of R100 and the Responder gets 20% of R100. If the proposal is rejected, both the Proposer and Responder earn nothing from this interaction. When the Responder hovers over a choice of Accept or Reject, the Responder is shown the amounts that the Proposer and the Responder would earn from this choice. In the screenshot, if the Responder accepts the proposed [80%, 20%] split, then the Proposer gets R80 and the Responder gets R20.



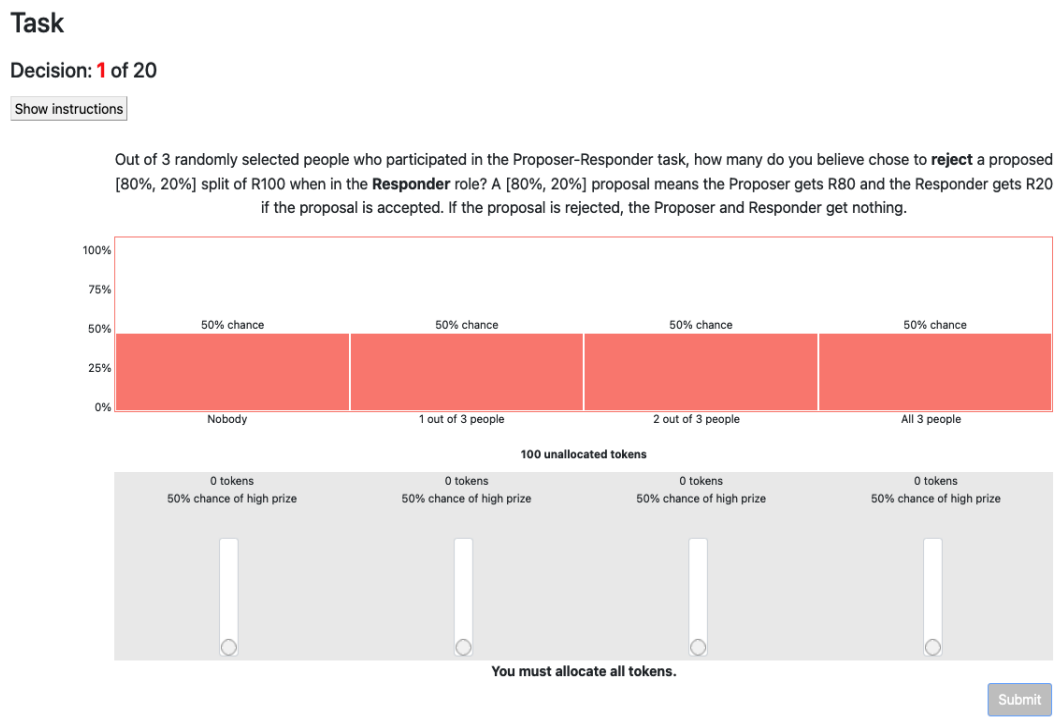
In this task you will be asked for your beliefs about the behaviour of others when they are in the *Responder* role, and also for your predictions of what other people in this study believe about these outcomes. There are four types of questions:

Type 1 - What do you believe UCT students **actually did** in the Responder role in response to the potential splits of [50%, 50%] and [80%, 20%]?
Type 2 - What do you predict the **other people** completing this task today believe about what UCT students **actually did** in the Responder role?
Type 3 - What do you believe UCT students in the Responder role **should have done** in response to the potential splits of [50%, 50%] and [80%, 20%]?
Type 4 - What do you predict the **other people** completing this task today believe about what UCT students **should have done** in the Responder role?

Question Type 1

The first type of question is about your beliefs concerning what UCT students **actually did** in the Responder role in response to the potential splits of [50%, 50%] and [80%, 20%]. For example, you will be asked "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [50%, 50%] split of R100 when in the **Responder** role?". You will also be asked "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". You will need to allocate 100 tokens to express your beliefs about the possible answers to each of these two questions.

This screenshot shows you what this task looks like for the case of an [80%, 20%] split.



You have 4 sliders to adjust, shown at the bottom of the screen, and you have 100 tokens to allocate across the sliders. Each slider allows you to allocate tokens to reflect your beliefs about the answer to this question. You must allocate all 100 tokens, and we always start with 0 tokens allocated to each slider. As you allocate

tokens, by adjusting sliders, the percentages displayed on the screen will change. Your potential earnings are based on the percentages that are displayed after you have allocated all 100 tokens, where higher percentages mean a higher chance of receiving a larger prize of R500 as opposed to a smaller prize of R50.

Where you position each slider depends on your beliefs about the correct answer to the question. The bars above each slider correspond to that particular slider. In our example, the tokens you allocate to each bar will naturally reflect your beliefs about the question, "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". The first bar corresponds to your belief that **Nobody** chose to reject a proposed [80%, 20%] split of R100. The second bar corresponds to your belief that **1 out of 3 people** chose to reject a proposed [80%, 20%] split of R100. The third bar corresponds to your belief that **2 out of 3 people** chose to reject a proposed [80%, 20%] split of R100. Finally, the fourth bar corresponds to your belief that **All 3 people** chose to reject a proposed [80%, 20%] split of R100. Each bar shows your percentage chance of earning R500 as opposed to R50, depending on what 3 randomly selected UCT students **actually did** in the Proposer-Responder task.

Let's look at an example to illustrate how you use these sliders. We'll imagine some actual numbers. These might not seem very likely to you; they are just for the sake of this example. Suppose you are answering the question we just discussed, "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". Suppose you think that out of 3 randomly selected UCT students who took part in the Proposer-Responder task, there is a good chance that "All 3 people" in the Responder role chose to reject a proposed [80%, 20%] split of R100. Then you might allocate 50 tokens with the slider for "All 3 people." Suppose you also think there is a pretty good chance that "2 out of 3 people" in the Responder role chose to reject a proposed [80%, 20%] split of R100. Then you might allocate 35 tokens with the slider for "2 out of 3 people." Finally, suppose you think there is a low chance that "1 out of 3 people" in the Responder role chose to reject this proposal, and an even lower chance that "Nobody" in the Responder role chose to reject this proposal. Then you might allocate 10 tokens with the slider for "1 out of 3 people," and 5 tokens with the slider for "Nobody."

This is what the display would look like if those were your choices:

In the example, because all 100 tokens have been allocated, the **Submit** button becomes clickable so that you can submit your choice and move on to the next question. If you would like to change your token allocation before clicking the Submit button, then just use the sliders to make any adjustments. For example, you could allocate zero tokens with one or more of the sliders.

Question Type 2

The second type of question is about what you predict the **other people completing this task today** believe about what UCT students **actually did** in the Responder role. In other words, you are predicting how the other people completing this task today actually allocated their 100 tokens for Question Type 1. Again, you will be asked to allocate 100 tokens to express your beliefs about the possible answers to these questions, but in this case you will have 10 sliders to adjust to reflect your beliefs about the answer to each question. Remember that for Question Type 1, exactly 100 tokens had to be allocated across the four sliders. So when allocating tokens for the next set of 4 questions, keep in mind that no subject could have allocated more or less than 100 tokens across the four sliders. This is what the set of questions look like:

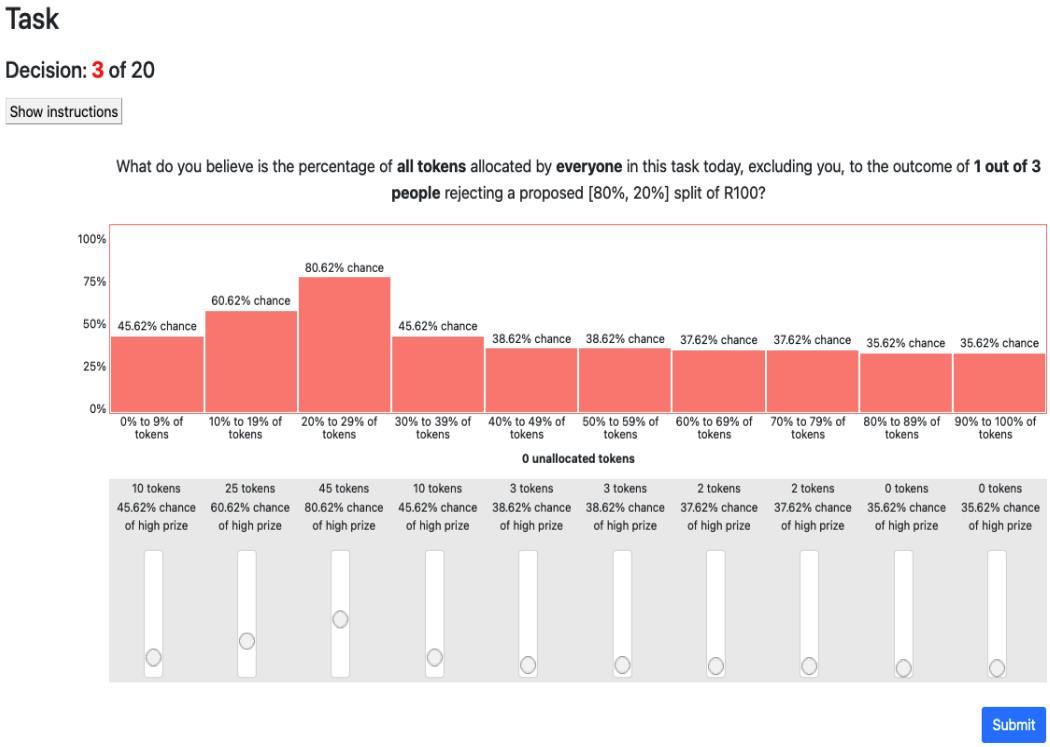
- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **Nobody** rejecting a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **1 out of 3 people** rejecting a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **2 out of 3 people** rejecting a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **All 3 people** rejecting a proposed [80%, 20%] split of R100?

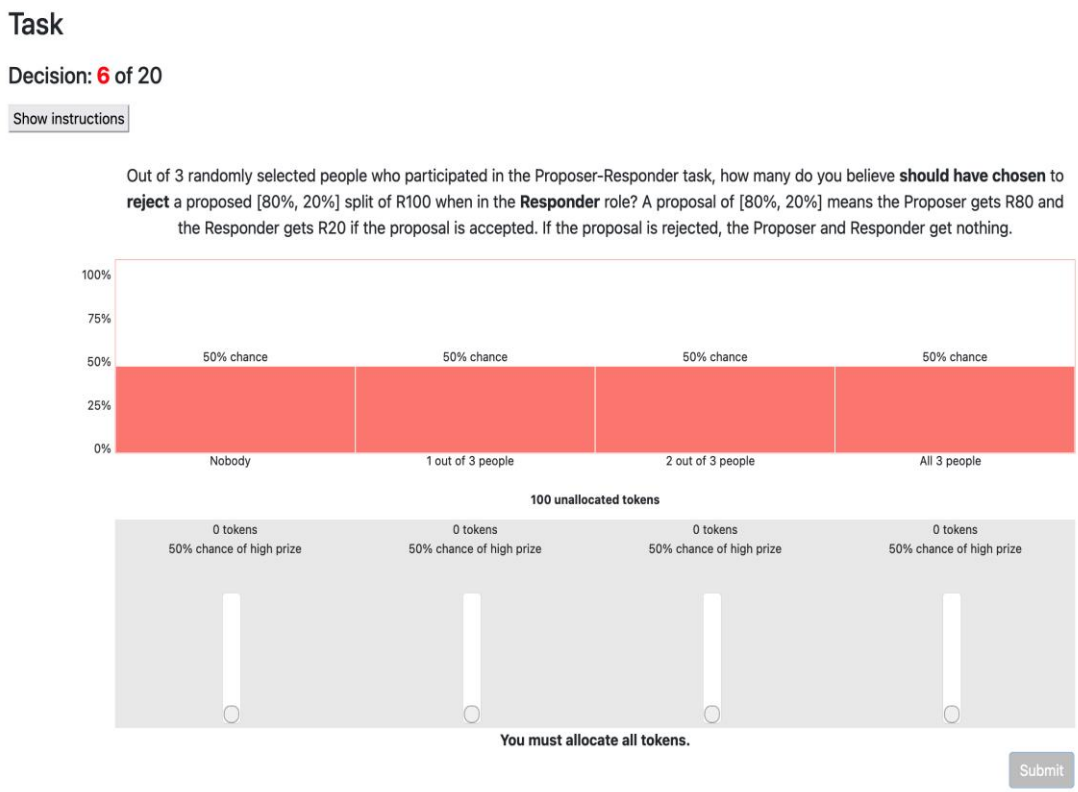
Let's look at another example. Once again we'll imagine some actual numbers that you might not think are very likely. But, as before, they are just for the sake of this example. Suppose you are answering the question, "What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **1 out of 3 people** rejecting a proposed [80%, 20%] split of R100?". Suppose that you predict there is a good chance that the people completing this task today allocated 20% to 29% of all tokens to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate 45 of your tokens with the slider for "20% to 29%." Perhaps you predict there is also a fairly good chance that the people completing this task today allocated 10% to 19% of all tokens to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate 25 tokens with the slider for "10% to 19%." Perhaps you think there is an equal chance that people allocated either 0% to 9% or 30% to 39% to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate 10 tokens with the slider for "0% to 9%," and 10 tokens to the slider for "30% to 39%." Finally, suppose you think there is a very low chance that more than 40% of all tokens were allocated to the outcome that **1 out of 3 people** rejected a proposed [80%, 20%] split of R100. Then you might allocate your remaining 10 tokens as shown in this screenshot.



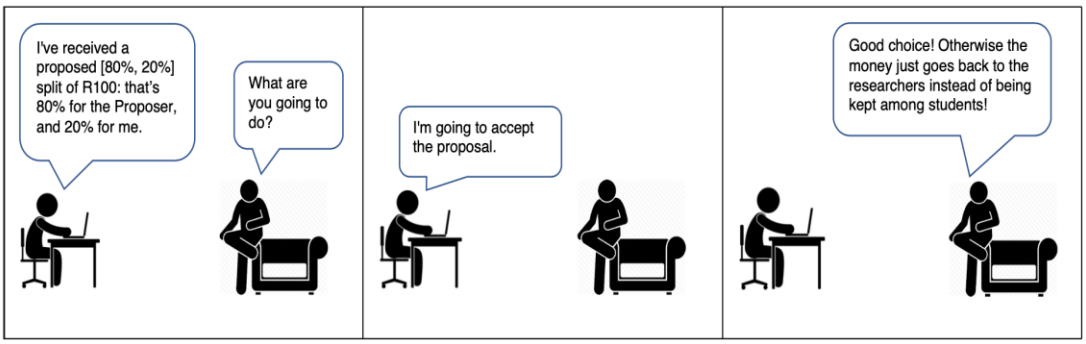
Remember you are free to allocate zero tokens with one or more of the sliders, just as we did for sliders 80% to 89%, and 90% to 100% in the screenshot.

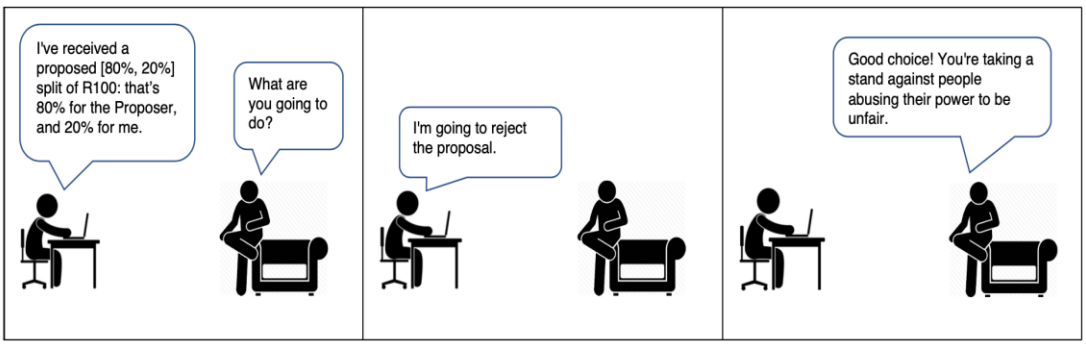
Question Type 3

The third type of question is about what you believe UCT students in the Responder role **should have done** in response to the potential splits of [50%, 50%] and [80%, 20%]. For example, you will be asked, “Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe **should have chosen** to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?”. You will need to allocate 100 tokens to express your beliefs about these two questions: one question for the potential [50%, 50%] split, and one question for the potential [80%, 20%] split. These questions will not be used to determine your payment for this task. However, the Type 4 questions, which *will* be used to determine your payment, relate to how other people allocated their tokens to the Type 3 questions, so please think carefully about your token allocation for the Type 3 questions. This screenshot shows you what these questions look like.



You might find it strange that someone could believe that different people **should** make different choices in this situation. These cartoons might help you to see how there could be variation in what people believe other people **should** choose.



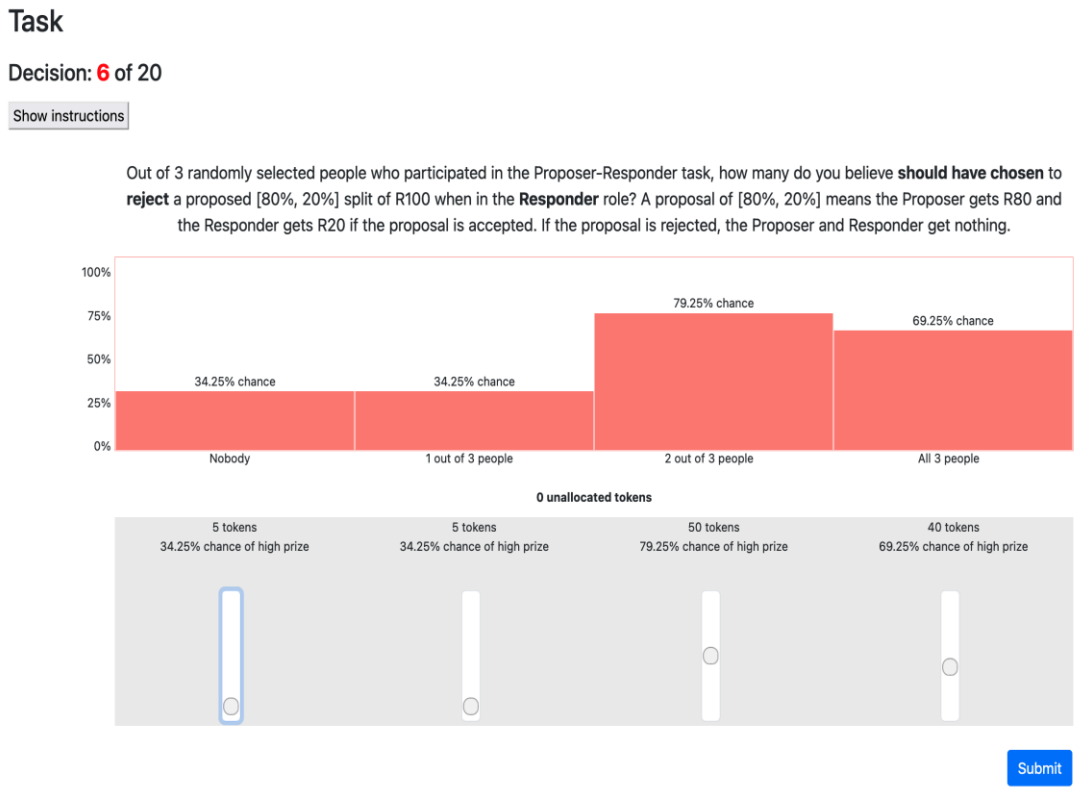


In the first cartoon, the person at the desk says, “I’ve received a proposed

[80%, 20%] split of R100: that's 80% for the Proposer, and 20% for me." The other person asks, "What are you going to do?". The first person responds by saying, "I'm going to **accept** the proposal." The other person says, "Good choice! Otherwise the money just goes back to the researchers instead of being kept among students!"

In the second cartoon, the person at the desk says, "I've received a proposed [80%, 20%] split of R100: that's 80% for the Proposer, and 20% for me." The other person asks, "What are you going to do?". The first person responds by saying, "I'm going to **reject** the proposal." The other person says, "Good choice! You're taking a stand against people abusing their power to be unfair."

Let's look at another example. Once again we'll imagine some actual numbers that you might not think are very likely. But, as before, they are just for the sake of this example. Suppose you are answering the question, "Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe **should have chosen** to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?". Suppose you think that out of 3 randomly selected UCT students who participated in the Proposer-Responder task, **at least** 2 out of 3 people **should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 50 tokens with the slider for "2 out of 3 people," and 40 tokens with the slider for "All 3 people." Suppose you think that the sliders for "Nobody" and "1 out of 3 people" are equally likely. Then you would allocate 5 tokens with the slider for "Nobody" and 5 tokens with the slider for "1 out of 3 people." This is what the display would look like if those were your choices:



Question Type 4

The fourth type of question is about what you predict the **other people completing this task today** believe about what UCT students **should have done** in the Responder role. In other words, you are predicting how the other people completing this task today actually allocated their 100 tokens for Question Type 3. Again, you will be asked to allocate 100 tokens to express your beliefs about the possible answers to these questions, but in this case you will have 10 sliders to adjust to reflect your beliefs about the answer to each question. Remember that for Question Type 3, exactly 100 tokens had to be allocated across the four sliders. So when allocating tokens for the next set of 4 questions, keep in mind that no subject could have allocated more or less than 100 tokens across the four sliders. This is what the set of questions look like:

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **Nobody should have chosen** to **reject** a proposed [80%, 20%] split of R100?

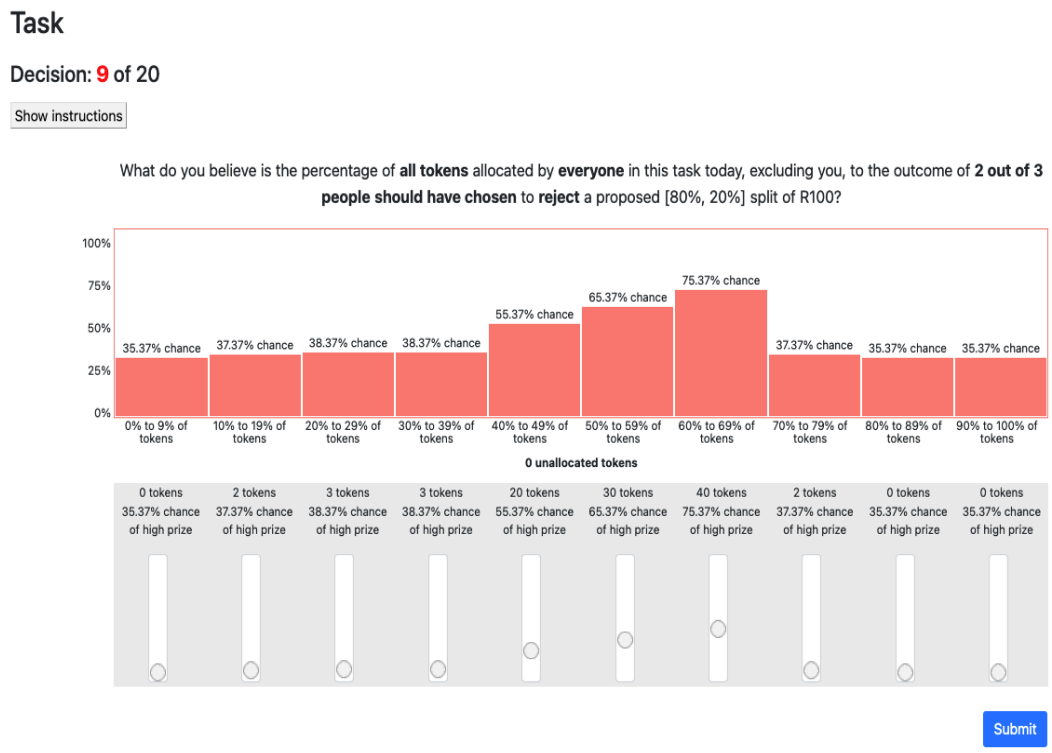
- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **1 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?

- What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **All 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?

Let's look at another example. Once again we'll imagine some actual numbers that you might not think are very likely. But, as before, they are just for the sake of this example. Suppose you are answering the question, "What do you believe is the percentage of **all tokens** allocated by **everyone** in this task today, excluding you, to the outcome of **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100?". Suppose that you predict there is a good chance that the people completing this task today allocated 60% to 69% of all tokens to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 40 of your tokens with the slider for "60% to 69%." Perhaps you predict there is also a pretty good chance that the people completing this task today allocated 50% to 59% of all tokens to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 30 tokens with the slider for "50% to 59%." Perhaps you think there is also a fairly good chance that people allocated 40% to 49% of all tokens to

the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate 20 tokens with the slider for "40% to 49%." Finally, suppose you think there is a very low chance that less than 40% or more than 70% of all tokens were allocated to the outcome that **2 out of 3 people should have chosen** to **reject** a proposed [80%, 20%] split of R100. Then you might allocate your remaining 10 tokens as shown in this screenshot.



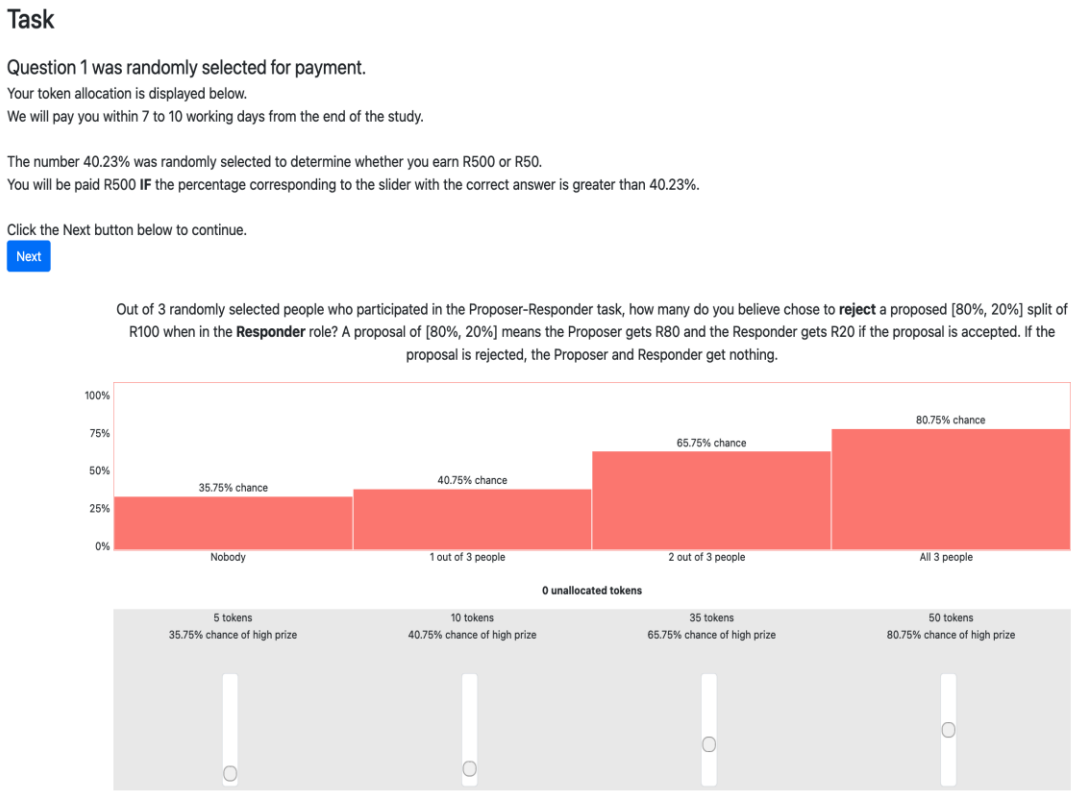
Task Earnings

You will be paid for one randomly chosen question from among your answers to the questions of type 1, 2 or 4. Your responses to that selected question will determine your earnings in this task as we explain now.

As mentioned earlier, by allocating tokens with the sliders the percentages on the screen change. These percentages are between 0 and 100, and they represent the probability of winning different money prizes depending on the allocation of your tokens. The prizes are either R500 or R50. The higher the percentage for a slider, the greater your chance of being paid R500 instead of R50. On the other hand, the lower the percentage for a slider, the smaller your chance of being paid R500 instead of R50. If you allocate all your tokens to one slider this gives you a 100% chance of being paid R500, if the correct answer is represented by that slider.

To determine payment for this task, the computer will randomly select one question of type 1, 2 or 4. The decision screen selected will be shown back to you and the computer will record the percentages you received from allocating your tokens. You will either be paid R500 or R50 depending on your token allocation and the correct answer to the randomly selected question.

For example, suppose a question of type 1 is randomly selected for payment.



And suppose the question that gets randomly selected is, “Out of 3 randomly selected people who participated in the Proposer-Responder task, how many do you believe chose to **reject** a proposed [80%, 20%] split of R100 when in the **Responder** role?”. Finally, suppose you allocated 35 tokens with the slider for “2 out of 3 people” and out of the 3 randomly selected UCT students, 2 of them actually did reject this proposal. With 35 tokens allocated to this answer you have a 65.75% chance of winning R500, and therefore a chance of 34.25% of winning R50, where 34.25% is just 100% - 65.75%. The computer will then randomly draw a number between 0 and 100, with every number between 0 and 100 being equally likely. If the randomly drawn number is less than or equal to the percentage you received for the correct answer, you will win R500. If the randomly drawn number is greater than the percentage you received, you will win R50.

Thus, your earnings depend on your reported beliefs and, of course, the true answer. For example, if the percentage you received is 65.75%, as in our example, and the number that is randomly drawn is 40.23%, you will be paid R500 since 40.23% is less than 65.75%. This randomly drawn number of 40.23% is shown in the screenshot. However, if the randomly drawn number is 79.71%, because this is greater than 65.75% you will be paid R50. Finally, if you allocate all of your tokens to one slider that corresponds to the correct answer this means you will be paid R500 with certainty, because a randomly drawn number between 0 and 100 will always be less than or equal to 100.

So it is up to you to balance the strength of your personal beliefs with the risk of them being wrong.

There are two important points for you to keep in mind when placing your bets.

1. Your payment depends on the accuracy of your beliefs about the interaction that took place between UCT students in the Proposer-Responder task, and on other people's beliefs that were expressed in this task today. So think carefully about the decisions you make.
2. More tokens allocated to the correct slider increase your chance of earning R500 and decrease your chance of earning R50. The percentage you receive for the correct answer will be compared with the randomly drawn number to determine whether you are paid R500 or R50.

You will be paid through Standard Bank Instant Money within 7-10 working days of completing the study.

Appendix C: Demographic Questionnaire

1. What is your current age?

[selected from drop-down list of values between 18 to 80]

2. Which of the following gender groups do you identify with?

[select one]

Female

Male

Other

Prefer not to Answer

3. Where are you currently taking part this survey? That is, what CITY and STATE/PROVINCE are you currently in?

4. What relationship status describes you currently?

[select one]

Single and never married

In a relationship, but not married

Married

Separated, divorced, or widowed

Other

5. In the last month, what was your total income from all sources?

6. On what day do you typically receive your income each month?

7. What is your current financial situation on the following scale?

[select one]

Very broke

Broke

Neither broke nor in good shape

In good shape

In very good shape

8. In what population group do you classify yourself?

[select one]

Black/African

Coloured

Indian

White
Prefer not to answer
Other

If other, please specify (optional)

9. How do you see yourself: are you a person who is fully prepared to take risks or do your try to avoid taking risks? Please select an option on the scale, where 0 means "not at all willing to take risks" and 10 means "very willing to take risks".

0 - Not at All Willing to Take Risks
1
2
3
4
5
6
7
8
9
10 - Very Willing to Take Risks

Appendix D: Normative Values Survey

Below are a series of alternative statements. Please tell us to what extent you agree with one statement or the other statement in each set. A score of 1 means you completely agree with the first statement and completely disagree with the second statement. On the other hand, a score of 10 means you completely agree with the second statement and completely disagree with the first statement. Scores between 2 to 9 mean you partly agree with both statements. Scores closer to 1 mean you agree more with the first statement and less with the second statement. Scores closer to 10 mean you agree more with the second statement and less with the first statement.

A. *ReligLaw (NOT SEEN BY SUBJECTS)*

1. Our religious beliefs should provide the basis for the laws of our country.
2.
3.
4.
5.
6.
7.
8.
9.
10. No single set of religious beliefs should be imposed on our country.

B. *EcoGrow (NOT SEEN BY SUBJECTS)*

1. We should protect the environment, and try to make our cities and countryside more beautiful.
2.
3.
4.
5.
6.
7.
8.
9.
10. We should encourage economic growth without environmental restrictions on business.

C. *ServTax (NOT SEEN BY SUBJECTS)*

1. We should improve government services and social assistance even if it means increasing taxes.
2.
3.

4.
5.
6.
7.
8.
9.
10. We should reduce taxes, even if it means reducing government services and social assistance.

D. Equallnd (NOT SEEN BY SUBJECTS)

1. There should be a more equal distribution of wealth.
2.
3.
4.
5.
6.
7.
8.
9.
10. There should be more incentives for individual initiative.

E. OrderLib (NOT SEEN BY SUBJECTS)

1. We should maintain law and order.
2.
3.
4.
5.
6.
7.
8.
9.
10. We should defend civil liberties.

F. PrivPub (NOT SEEN BY SUBJECTS)

1. We should privatise public enterprises.
2.
3.
4.
5.
6.
7.
8.
9.

10. We should maintain existing public enterprises.

G. Abortion (NOT SEEN BY SUBJECTS)

1. Abortion should always be illegal.
2.
3.
4.
5.
6.
7.
8.
9.
10. Abortion should remain legal.

H. Compete (NOT SEEN BY SUBJECTS)

1. Work hard and compete, so that you can get ahead at work.
2.
3.
4.
5.
6.
7.
8.
9.
10. Aviod competition with fellow workers in order to maintain good relations.

I. Particip (NOT SEEN BY SUBJECTS)

1. Increase citizen participation in government decision making.
2.
3.
4.
5.
6.
7.
8.
9.
10. Government should quickly make decisions based on the knowledge of experts.

J. Community (NOT SEEN BY SUBJECTS)

1. Each person should put the well-being of the community ahead of their own interests.
2.

3.
4.
5.
6.
7.
8.
9.
10. Everybody should be free to pursue what is best for themselves as individuals.

K. OurWay (NOT SEEN BY SUBJECTS)

1. Our country should defend our way of life instead of becoming more and more like other countries.
2.
3.
4.
5.
6.
7.
8.
9.
10. It is a good idea to copy practices from other people all over the world.

L. GovParent (NOT SEEN BY SUBJECTS)

1. People are like children; the government should take care of them like a parent.
2.
3.
4.
5.
6.
7.
8.
9.
10. Government is an employee; the people are the bosses who control the government.

M. Conflict (NOT SEEN BY SUBJECTS)

1. Conflict should be avoided at all costs in our society.
2.
3.
4.
5.

6.
7.
8.
9.
10. Conflict is a normal part of a society.

N. TimeResolves (NOT SEEN BY SUBJECTS)

1. Most problems can be resolved with time.
2.
3.
4.
5.
6.
7.
8.
9.
10. When confronted with problems, we should address them as soon as we can.

O. GovWellBeing (NOT SEEN BY SUBJECTS)

1. The government should bear the main responsibility for the wellbeing of people.
2.
3.
4.
5.
6.
7.
8.
9.
10. People should look after themselves and be responsible for their own success in life.

P. Customs (NOT SEEN BY SUBJECTS)

1. It is better for society if different racial and ethnic groups maintain their distinct customs and traditions.
2.
3.
4.
5.
6.
7.

8.
9.
10. It is better if these groups adapt and blend into the large society.

Q. COVID (NOT SEEN BY SUBJECTS)

1. The government should devote many more resources to combating COVID, even if this means that less money is spent on things like education.
2.
3.
4.
5.
6.
7.
8.
9.
10. There are many other problems facing this country beside COVID; even if people are dying in large numbers, the government needs to keep its focus on solving other problems.

The survey has now ended. Thank you for your participation.