DOES DESIRABILITY BIAS IMPACT AMBIGUITY ATTITUDES?

AN EXPLORATORY STUDY

G.K. KONDOV

(ID 466530)

Erasmus School of Economics, Erasmus University Rotterdam

MSc. Business and Economics – Behavioural Economics

Supervisor: Prof. Dr. Peter P. Wakker

TABLE OF CONTENTS

Abstract

"*The main goal of the present study is to investigate the possible role of desirability bias on individuals' decisions regarding uncertain future events. This is achieved by operationalizing two key constructs – event (un)desirability and ambiguity attitudes - through an economic experiment. The empirical evidence generated by this study is inconclusive with regards to the suggested tentative hypotheses. Hence, no systematic association, nor evidence for causation, was found between the main variables of interest. I carefully consider several possible explanations for this result. However, this exploratory study presents a novel direction for further research into the components of ambiguity attitudes.*"

1. INTRODUCTION (DESIRABILITY BIAS AS A POSSIBLE COMPONENT OF AMBIGUITY ATTITUDES)

"*People can foresee the future when it coincides with their wishes*." – George Orwell

People are usually faced with an abundance of external and internal stimuli begging for their attention, to a degree far outstripping our total momentary sensory capacity. Because of our inability to process, or indeed even attend to, such huge quantities of information, we have developed mental shortcuts we rely on to search for, attend to, and interpret information. Therefore, we are constantly making conscious and unconscious choices regarding which stimuli to ignore or filter. The outcome from this process is highly subjective and personal but has enormous implications for our choices, attitudes, behaviors, and expectations, among others.

Our brains rely on heuristics and habits to manage our limited mental resources in the most efficient ways by automating and streamlining many decisions in different areas of our lives. In his best-selling book, Kahneman (2011) calls this "System 1" decision-making. Unfortunately, such strategies often lead to predictable biases in our judgements. One such area that is impacted and that has drawn a significant amount of research is judgement about likelihood of future events. As the quote at the beginning from British writer George Orwell implies, there tends to be a consistent and predictable relationship between desires and predictions about the future. As pointed out by Lench (2009), many studies (predominantly in psychology) have demonstrated this relationship in field experiments (and to a lesser degree in the lab) but very few have been able to provide convincing empirical proof of a causal effect.

Apart from research psychologists, decisions about future events and what influences them have also occupied the minds of academic economists for the past decades. Since the time the formal conceptual distinction was first made between the theory of risk and that of uncertainty, many research studies have been published investigating how we make decisions about future events when the probabilities are unknown or unknowable. In contrast to their colleagues in psychology, economists have preferred the lab over the field for their experiments in this area due to the difficulty of keeping subjective likelihood evaluations fixed when examining natural events. This artificial setting for the study of decisions under uncertainty resulted in the well-known empirical finding of preference for risk over ambiguity, termed ambiguity aversion. The empirical status quo remained unchallenged for a long period until Heath and Tversky (1991) published their famous paper, positing that ambiguity aversion should be thought of as a special case of a broader set of attitudes towards ambiguity. By using natural events instead of the classic Ellsberg's urn problem, they demonstrated the all-important ambiguity characteristic of "*source-dependence*" (where source refers to the mechanism generating the uncertainty).

The goal of the present study is therefore two-fold. First, to contribute to the empirical literature documenting that a wider set of ambiguity attitudes are possible when examining natural events instead of artificial. Second, to investigate the role of desirability bias on our decisions regarding uncertain future events. This study also tests the effects of undesirable events on decision-making under ambiguity, also including desire-neutral stimuli. As pointed out in Lench (2009), past studies manipulating the desirability of events rarely include a neutral condition and they are unconvincing when it comes to distinguishing whether it is the "*desirable*" or "*undesirable*" events which drive the desirability bias. This study could be one of the few that was able to shed light on such questions.

The structure is as follows. Section 2 reviews the theory of desirability bias and briefly introduces the concept of ambiguity attitudes. Section 3 details the economic experiment, outlining the key methods and design decisions. I wrap up the section with a few tentative hypotheses based on knowledge from extant theoretical and empirical research. Those are only tentative due to the exploratory character of this study. In section 4, I present the formal plan for analyzing the data and testing the hypotheses. Section 5, then, shows the graphical and numerical output from the statistical analysis conducted with the help of Stata. Section 6 concludes by discussing the key lessons learned from this experiment as well as the limitations of the study.

2. LITERATURE REVIEW & THEORETICAL BACKGROUND

This section is split into several parts. Subsection 2.1 introduces the idea of "*ambiguity attitudes*", its intellectual origins and real-world significance. Subsection 2.2 introduces the concept "*desirability bias*" and further delves into the relevant theory and empirics. Because of the large area covered by this subsection it is further split into four parts – 2.2.1 distinguishes desirability bias from similar psychological decision-making phenomena, 2.2.2 attempts to uncover through theory the underlying psychological sources of the bias, 2.2.3 explores different mechanisms through which desirability bias may act to distort decision-making, and 2.2.4 closes with an in-depth review of the empirical literature.

2.1. AMBIGUITY ATTITUDES

The formalization of the theory of decisions under uncertainty (and hence its separation from the theory of decisions under risk) is often credited to Knight (1921) who it is believed was the first to make the conceptual distinction between measurable and unmeasurable uncertainty. Subsequently, Daniel Ellsberg (1961) famously placed this issue firmly in the spotlight of economic theory with his two-color urn problem. Since his paper, people's preference for clear over vague bets has become one of the stylized facts in the field of decisions under uncertainty (e.g. see Camerer and Weber, 1992). His findings, however, represented a major challenge to the established economic orthodoxy at that time because they painted a picture of a severely limited relevance of the central model for analyzing such decisions – subjective expected utility. As observed in Heath and Tversky (1991), this is because, naturally, most decisions in practice involve unmeasurable uncertainty in contrast to what dictates the events in a game of chance. Heath and Tversky felt compelled to explore the usefulness of ambiguity theory in the real world and were, therefore, among the first to pose the question whether the widely reported ambiguity aversion phenomenon would extend beyond the domain of games of chance (such as the Ellsberg's urn problem) and of stated probabilities. What they find confirms their hypothesis that a further component needs to be added to ambiguity – that of perceived level of knowledge of the subject matter. They find that people prefer betting on high-perceived-competence vague events over chance events but prefer the latter over low-perceived-competence vague events. Hence, a key exception to ambiguity aversion was established.

It was also this realization of the far-reaching implications and real-world relevance of this area of economics that prompted Fox and Tversky to look with a critical eye at the body of empirical evidence accumulated at that point and attempt to explain the findings of Heath and Tversky (1991). They extended the line of empirical research on ambiguity that uses the setting of naturalistic (or natural) events - those encountered in everyday life, such as "*[t]he decisions to undertake a business venture, to go to court, or to undergo medical treatment…*" (p. 586) – and discovered further exceptions to the ambiguity effect. Their realization that virtually all studies that preceded theirs employed a within-subject design enabled them to find that ambiguity aversion is also eliminated if vague and chance events are evaluated in isolation, i.e. as in the case of between-subject experiments. These two studies are the early standouts in a series of papers which later followed that treat ambiguity aversion as a special case of source preference, i.e. holding subjective probabilities and outcomes of ambiguous events fixed, people prefer (i.e. demonstrate a reduced or altogether eliminated degree of ambiguity aversion) some sources of uncertainty over others (Trautmann and de Kuilen, 2014). With that in mind, Trautman and de Kuilen coin the term "*ambiguity attitudes*" to refer to people's behavior when faced with unmeasurable uncertainty and to reflect the greater depth of the concept.

Apart from the competence and comparative ignorance component, another popular factor in ambiguity attitudes is the so-called home bias or familiarity effect. There is empirical evidence (e.g. Chew et al. 2008 and 2012, Abdellaoui et al. 2011) to suggest that the ambiguity effect is also reduced or even reversed for sources associated with high degrees of perceived familiarity (not to be confused with knowledge which can be viewed as a stronger form of familiarity). In this paper, I explore the usefulness of a further potential component of ambiguity attitudes – desirability bias - and similarly to the recent trend in the experimental literature on decisions under uncertainty also test this with natural sources. Desirability bias is a well-researched concept in the psychological literature, yet it has received virtually zero attention so far from academic economists.

2.2. DESIRABILITY BIAS - WHAT IS IT?

As observed by Kunda (1990), the possibility of motivations or goals (here, equivalent to desires and wishes) influencing cognition is not a novel concept in psychology. According to this strand of research, motivations play a cognitively mediating role and generally fall into one of two categories. First, there are motivated reasoning phenomena which make us strive to be accurate and precise in our predictions. Second, there are also such motives that encourage us to reason our way into a desired conclusion. Kunda proposes that motivations influence reasoning task outcomes by biasing various cognitive processes such as those related to the active recollection, formation and evaluation of convictions. Therefore, the two general categories of motivating phenomena impact in distinct ways the choice set of strategies used for the active recollection, formation and evaluation of convictions. These distinct ways are dictated by the reasoning task outcomes the individual strives towards.

In other words, motivated reasoning phenomena encouraging precision would make us select the most appropriate strategies given the specificities of each reasoning task. In contrast, motivated reasoning phenomena encouraging the arrival at a desired conclusion would make us select those strategies which facilitate reaching the desired conclusion. The author also concedes the possibility of motivated reasoning also acting as a behavioral mediator (for example, see Stuart et al, 2017, sub-section 2.2.3 of this paper).

Desirability bias, a phenomenon related to the broader concept of motivated reasoning (see subsection 2.2.1 for a discussion of similarities and disparities) and also referred to as "*wishful thinking*" or "*value bias*", can therefore have both cognitively and behaviorally regulating roles. It is commonly defined as the tendency to over-estimate the likelihood of subjectively desirable events and under-estimate the likelihood of subjectively undesirable events (e.g. Budescu and Bruderman, 1995). However, arguably, the effect of this bias on decision-makers is much more nuanced than simply the distortion of probability judgements. I opine that a more suitable way to define desirability bias would be to take a broader view and consider the systematic relationship between the subjective appeal of an event and the choices, judgement or behavior towards this event such as but not limited to over- or under-estimating its likelihood (e.g. Massey, Simmons, and Armor, 2012), over- or under-preparing for it (e.g. Stuart et al., 2017), over- or under-evaluating the strength of the evidence favoring or disfavoring it (e.g. Bar-Hillel and Budescu, 1995), or making asymmetric, subjectively arbitrary bets associated with it (e.g. Windschitl et al., 2010).

Bar-Hillel and Budescu (1995) originally coined the term "*desirability effect*", conceptualized and tested it as such in an experimental study exploring the link between event desirability and inflated perceived likelihood. However, they do concede that event desirability may not just express itself via biased evaluation of probability and later in this chapter I will present arguments as to why this initial definition could indeed be considered too limited and explore several other possible channels through which the subjective desire associated with an event may influence our attitudes.

Lench (2009) also observes the lack of agreed-upon definition in the academic literature so far regarding what constitutes "*desirability*" and the resulting difficulty in determining the causal influence on and consequences of event desirability for decision-making. Most often, she continues, event desirability or "*value*" is treated as a permanent or constant phenomenon, whereas she reminds that a significant body of literature has shown that motivation and affective states influence the evaluation of stimuli. Hence, desirability bias can be explained in terms of transient reactions and captures attitudes much richer and inherently dynamic, if not ephemeral. In her treatment of desirability, the main mechanism affecting the evaluation of stimuli is fast, difficult to control, and susceptible to incidental affect.

The next four sub-sections (2.2.1 to 2.2.4) dig deeper into the concept and practical significance of desirability bias by distinguishing it from similar response biases (2.2.1), exploring the underlying psychological drivers of this bias (2.2.2), investigating several possible channels through which the bias influences our interaction with the world around us (2.2.3), and reviewing the relevant empirical literature on its existence and magnitude as found in laboratory and (quasi-)field experimental studies (2.2.4).

2.2.1. DISTINGUISHING FEATURES (HOW IS IT DIFFERENT?)

To pin down what exactly constitutes desirability bias, Krizan and Windschitl (2007) set about distinguishing it from other related concepts – namely, motivated reasoning, over-optimism, and the preference-expectation link.

First, they argue that motivated reasoning can be applied to all sorts of mental tasks, such as the evaluation of the strength of an argument or likelihood associated with a personally relevant event. In contrast, wishful thinking, by their definition, applies exclusively to predictions. A key implication of this difference is that predictions unlike many other forms of judgement are sooner or later going to be subjected to what is colloquially known as a "reality check" or what the authors call a "*moment of* *truth*". This expectation that they will eventually learn whether their forecasts came true or not may influence decision-makers in a way strictly distinct from what occurs with other types of judgements and evaluations. For example, manifestation of desirability bias, such as inflated optimism or tending to guess in a favorable direction (more on these mechanisms in 2.2.3), may be suppressed due to a strong aversion to potentially experiencing disappointment. Such behavior has been termed "*bracing for loss*" (e.g. see van Dijk, Zeelenberg, and van der Pligt, 2003). Another reason for the absence of strong desirability bias effects may be superstition or conviction that one can "*tempt fate*" or "*jinx*" the realization of the desired event if behaving in a way consistent with one's own desires. These psychological phenomena are notable examples of the multifaceted manner in which desirability may impact behavior or judgement, and, perhaps surprisingly, illustrate why desirability bias would sometimes result in behavior or judgement which appears to act against one's own wishes.

Second, the previous paragraph also helps draw a line between wishful thinking and over-optimism. It is argued that the presence of wishful thinking does not necessarily bring about over-optimism, in fact quite the contrary in some instances. Krizan and Windschitl (2009) also point out that observing over-optimistic behavior or decisions also does not necessarily imply that wishful thinking is behind this. They suggest that there could be other underlying psychological forces such as cognitive egocentrism, being insufficiently informed, and "*gambler's fallacy*". Bar-Hillel and Budescu (1995) also hold the view that wishful thinking is closely related to but distinct from optimism and hopefulness. However, they do not elaborate further on this notion.

In addition, the authors distinguish desirability from optimism bias based on whether the decision-maker can influence the outcome of an ambiguous event (see also Vosgerau, 2010). They classify optimism bias as relevant for events over which the decision-maker has control and that it involves a host of issues such as efficacy, commitment and effort which are not associated with desirability bias. In contrast, the established definition of desirability bias associates it with events over which the decision-maker has no control and there are no steps he or she can conceive of to influence the resolution of an ambiguous event. In that sense, they depart from Bar-Hillel and Budescu (1995) who also make the conceptual distinction between self-centered (the feeling that one can "*defy the odds*", formally termed above-average and comparative optimism effects in Chambers and Windschitl, 2004) and "*world at large*" (general rose-colored perspective on the world) events but define desirability effects as relevant for both types of events. Stuart et al. (2017) suggest other influences on behavior in cases of ambiguous events over which the decision-maker has no control – superstition (e.g. fear of tempting fate, Risen and Gilovich, 2008). A related idea is that of "*illusion of control*" (e.g. desired outcome of a coin toss more likely if the subjects themselves toss the coin).

To summarize up until here, the interpretation followed in this paper is that over-optimism or optimism bias appears to be an all-encompassing umbrella term, capturing many different underlying mechanisms. One such mechanism may be desirability bias. Hence, optimism-biased choices do not necessarily imply desirability bias. While the latter does not necessarily influence choices solely by biasing the perceived likelihood of events of different desirability levels.

Third, Krizan and Windschitl (2007) believe that another popular result such as the preference-expectation link also does not offer unequivocal proof for the presence of desirability bias. It is argued, rightfully, in my humble opinion, that there could be several plausible explanations for the established correlation between people's wishes and their stated predictions, such as omitted variable bias (e.g. social influence or competence) or reverse causality (e.g. the "*bandwagon effect*" or problem of causal flow, e.g. Granberg and Brent, 1983 p. 483). This makes it difficult to argue persuasively for a direct cause-and-effect explanation in the desired direction. Krizan, Miller, and Johar (2010) make a similar argument. They provide different examples from the domain of politics, illustrating how the causal effect could be running in the opposite direction. First, preferences may follow expectations (rather than the other way around) if voters change their preferences to vote for the candidate with the greatest likelihood of winning. The opposite could also be true if voters would rather support the "dark horse" candidate instead. Finally, they also concede the possibility of expectations influencing preferences subconsciously. All these problems again stem from casual use of the concept of causality and unconvincing experimental evidence.

Budescu and Bruderman (1995) distinguish desirability bias from the illusion of control effect as two very different response biases which are only rarely affected by identical factors. They define the latter as "*the belief that one can manipulate, influence, and control outcomes of chance events*" (p.109). They briefly survey the past literature addressing the relationship between the two biases. More broadly, the relatively few previous studies on this topic model the two biases as conceptually confounded and having a positive relationship (e.g. Koehler et al., 1994), having a causal relationship running from illusion of control to desirability (e.g. Wagenaar and Keren, 1988), view the illusion of control bias as a necessary condition for the occurrence of desirability bias (e.g. McKenna, 1993), and so forth. In their own experiment, they find consistent evidence for desirability effects irrespective of the perceived level of control, showing that the desirability bias can indeed exist independently from the illusion of control bias. The illusion of control bias is found to be much weaker, mainly relevant for one-off tasks (apparently it virtually disappears in repeated tasks), and largely uncorrelated with the magnitude of the desirability bias.

Furthermore, Tappin and van der Leer (2017) examine desirability bias as a potential way of understanding how people incorporate new information and contrast it to confirmation bias. According to this take, desirability bias distorts the way new information is weighted, or even taken into consideration at all, depending on whether this new information would be mentally labeled as desirable or not. On the other hand, confirmation bias serves pre-existing convictions and makes individuals less likely to incorporate new information that challenges or disproves their opinions. However, the authors concede the difficulty of empirically separating the two effects and argue that their experimental study is one of the few to achieve this feat. They design their experiment to take advantage of electoral attitudes prior the 2016 US Presidential Election, and, in particular, the rare occurrence of conflict between preferences and expectations among voters. Their findings suggest that indeed new information is more likely to be incorporated if it supports a desirable resolution, irrespective of the prior subjective evaluation of the likelihood of this resolution (i.e. even in the case when the desirable information acts to disprove existing convictions). Hence, the desirability bias appears to dominate confirmation bias when the two biases go in opposite directions.

2.2.2. UNDERLYING PSYCHOLOGICAL ORIGINS OR THEORIES OF DESIRABILITY BIAS (WHERE DOES IT COME FROM OR WHAT CAUSES IT?)

Mayraz (2011) investigates different theories of desirability bias by outlining a set of conditions or circumstances under which desirability bias is likely to be influential. The three main theories he considers are the universal theory of wishful thinking, ego-utility theory, and cognitive bias theory. The first theory states that there is no restriction on the type of subjective judgements which are potentially sensitive to desirability bias. People behaving according to this theory can still be considered rational, assuming that the existence and intensity of desirability bias is arrived at after mentally trading-off the benefits of engaging in wishful thinking (e.g. see Taylor and Brown (1988) on positive illusions and mental health) to the drawbacks of potentially making sub-optimal judgements (e.g. experiencing disappointment). On this topic, Massey, Simmons, and Armor (2011) suggest and refer to studies (p. 7) presenting this issue as a cost-benefit analysis or examining it from the perspective of the "*optimal margin of illusion*".

Mayraz underlines that this theory implies larger desirability bias in cases when precision incentives are weak compared to a task where they are strong (empirical evidence on this conjecture is presented in 2.2.4). This variant of the universal theory is known as "strategic". In contrast, the "non-strategic" variant treats desirability effects as bundled with subjective evaluations and makes the prediction that wishful thinking would be present even in high-stake situations. According to ego-utility theory (see Koezsegi, 2006) wishful thinking effects are not universal but limited to self-centered, personally-relevant events, while the third theory disputes the very existence of desirability effects and attributes the purported evidence for wishful thinking to purely cognitive biases (Weinstein and Lachendro, 1982, on egocentrism).

After observing that expectations about US presidential elections were tightly linked to subjective preferences in the 1952-1980 period, Granberg and Brent (1983) also consider three possible explanations for their findings based on existing theories. They take link between preferences and expectation to imply the existence of wishful thinking effects, and hence the theories they propose are essentially theories of the origination of desirability bias. First, they point to work done by Heider in the 1940s and 1950s on the so called "*Balance Theory*". It describes a hypothetical situation in which a contradiction may arise between one's cognition and feelings (or wishes) which may then lead to actions taken by the conflicted person to "*remedy*" this internal inconsistency. They illustrate this theory with an example from the political domain. If a voter has a favorable attitude towards the general electorate but disagrees with this body of people on the most suitable presidential candidate, then an internal inconsistency is hypothesized to arise. According to Heider, the supposedly conflicted person can then "*rectify"* this inconsistency by altering their attitude towards the electorate, the candidate, or their estimate of the most likely behavior of the former. Granberg and Brent also believe, similarly to Bar-Hillel and Budescu (1995), that their findings can be treated as a special case of and interpreted according to the "*Pollyanna Principle*". This principle states that people are subconsciously wired to dwell on positives and filter out the negative, although some may consciously identify themselves as pessimists. The third and final theory they consider comes from the work of Goffman (1959) on the concept of impression management. This concept concerns the assumed intuitive understanding of people that personal traits such as optimism and confidence as well as those who possess them are likely to be evaluated more favorably than pessimism and insecurity. The self-presentation efforts of people can then turn wishful thinking into a habit, and hence their outer expressions such as predictions or choices can be misleading and not representative of their true attitudes. Uhlaner and Grofman (1986) write about the innate irrationality hypothesis, concluding that, taken together, these three explanations imply that wishful thinking is a congenital psychological trait for humans. In this context, Granberg and Brent connect these behaviors to the theory of evolutionary significance.

The final theoretical explanation I am going to consider is the one by Lench (2009), who investigates the possibility that immediate affective reactions which serve to mark an event as desirable (or not) may be an important underlying force behind or component of desire (as in "*event desirability*"). Furthermore, they may also be causally linked to ("*sufficient and necessary*" in the author's words) desirability bias, and as such Lench would be "…*first to establish that affective reactions to future events are the driving force behind the desirability bias*" (p. 197), in the context of future event likelihood evaluations. Hence, her arguments are limited to the impact of emotional affect on desirability bias in probability judgements only and ignores other possible mechanisms (discussed in detail in 2.2.3). Her approach is grounded in the intuitive assumption that affect is linked to event desirability (making the latter a dynamic event attribute and perhaps even unpredictable) and is explicitly distanced from the strand of economic theory which posits that desirability is a relatively more fixed attribute compared to other concepts (both between- and within subjects, i.e. the value of an event is consistent between people and across time periods).

She explores two conceptual accounts (theories establishing the direction of the effect) of the link between the desirability bias in probability judgements and affective responses. The first strand of this academic literature is represented by Chambers & Windschitl (2004) who instead seek to explain behavioral phenomena such as the above-average effect and the comparative-optimism effect not through desirability bias for example, but through nonmotivated (self-centered events) accounts such as information-processing limitations or as (Unintended?) consequences of experimental design and methodology choices. This alternative conceptual account is in line with the artifactual account of desirability bias proposed in Windschitl et al. (2010) as one of four possible channels through which event desirability influences choices, judgements, behaviors etc. (refer to section 2.2.3 for detailed description of this account). The second conceptual account she explores places a higher weight on negative affective reactions compared to positive, suggesting that the desirability bias in probability judgements may be driven by perceived negative event desirability. A similar prediction is made by the conceptual account proposed in Loewenstein et al. (2001). They also consider the role of affect at the exact moment a decision is being made and emphasize the role of emotions in their "*risk-as-feelings*" hypothesis of decision-making under risk or uncertainty. According to these interpretations, negative affective responses to undesirable events will provoke avoidance behavior and an exaggerated perception about the likelihood of these threats to the emotional well-being of the individual. In contrast, Lench predicts that the desirability bias caused by negative affect from undesirable events will lead to the underestimation of the likelihood of these undesirable events.

2.2.3. MECHANISMS (HOW DOES IT MANIFEST ITSELF?)

The term "*desirability effect*" was originally coined in Bar-Hillel and Budescu (1995) to refer to the link between event desirability and inflated perceived likelihood. However, they also consider a broader set of manifestations of wishful thinking such as action, attitude, motivation, belief. Since then, others have expanded the list of possible mechanisms through which the desirability effect impacts our daily lives.

As observed in Windschitl et al. (2010), few studies (exception is Budescu and Bruderman, 1995) attempt to uncover the different channels through which desirability effects creep into judgements and forecast, and potentially help explain why desirability effects appear stronger in some research paradigms but not in other (the different research paradigms are discussed at length in 2.2.4). They describe four psychological mechanisms driving the oft-discovered desirability bias in marked-card research paradigms and explore what these different mechanisms imply for the existence and magnitude of desirability effects in other experimental research paradigms.

First, as mentioned in the previous paragraph (and in section 2.2.2), the authors consider the possibility of experimenter bias driving the desirability effect findings in previous studies. Hence, one of the four psychological mechanisms, in fact, suggests that observed results are being misattributed to the influence of desirability bias. They suggest that many authors may prematurely assign desirability bias as the leading cause of the results they obtain, referring to this as the artifactual account of desirability bias (see Rosenthal and Fode, 1963). In the studies they single out, the experimenter would explicitly communicate the monetary value (may be a gain or a loss for the participant) of drawing a marked card and ask the participant to say out loud for his or her prediction. The authors find it plausible that participants may feel under pressure to answer in a specific way and be influenced by the framing of the question itself (e.g. asking whether a marked card will be drawn or asking whether a non-marked card will be drawn). Second, desirability effects are theorized to be driven by the manner in which the evidence for every possible event or resolution is perceived and evaluated, called the "biased evaluation" account. The authors point to an abundance of empirical studies from the related motivated reasoning literature showing that evidence is treated differently depending on whether it contradicts or supports a desired conclusion (e.g. see Balcetis, 2008 and Roese & Olson, 2007).

The final two mechanisms include the "biased threshold" and the "biased guessing" accounts, which assume unbiased likelihood evaluation but biased decisions or predictions due to either a lowered threshold for desirable events compared to undesirable events (and vice-versa, see Dawson, Gilovich, and Regan, 2002) or optimistic guessing. In the former case, a biased decision threshold may lead to different predictions for two events which are otherwise perceived to be equi-probable but have different degrees of (subjective) desirability. The example that they give is one where a marked card with a constant subjective probability of being drawn from a deck of cards across different experiments triggers a positive prediction more often in experiments when this event is desirable compared to another when it is undesirable. Regarding the biased-guessing account, they posit that it represents an asymmetric manner of providing an arbitrary answer to a dichotomous task. In other words, when decision-makers must resort to providing a "*subjectively arbitrary*" answer or guessing, they will tend to guess in an optimistic rather than pessimistic direction, i.e. favoring their desired conclusion (Stuart et al. 2017). This account is theorized to be most powerful when there is no perceived imbalance of evidence or knowledge leaning towards a specific event or resolution (effect colloquially known as "hunch" or "gut feeling").

As mentioned in sub-section 2.2, Kunda (1990) offers a comprehensive analysis of motivated reasoning (loosely using concepts such as goals or motives to mean any wish, desire or preference for a specific resolution) and explores different mechanisms for motivated directional biases. The key proposition here is that while people generally tend to arrive at the desired conclusions, they do not do so in an unconstrained manner. In other words, in order to reach the desired outcome of a reasoning task, the individual must first come up with enough seemingly reasonable justifications for that. These mechanisms are the essence of the selective use of strategies related to the recollection, formation and evaluation of convictions, such that enable the individual to muster up this evidence. The author proposes several such mechanisms which bias the use of strategies for solving the cognitive task at hand, such as biased accessing of beliefs (including dissonance research, biased beliefs about events and others), biases in selection of statistical heuristics, and biased research evaluation.

On the conjecture of desire affecting behavior, Stuart et al. (2017) suggest another possible manifestation of desirability bias – anticipatory or preparatory actions and choices. They consider two possible consequences for a decision-maker's tendency to act in the case of desirability bias. On the one hand, event desirability might lead to an increased tendency to act like the desired event will happen, or vice-versa that the undesired event will not happen. On the other hand, event desirability may also lead to a decreased tendency to act like the desired event will materialize, and vice-versa. The reader may recall that one potential explanation for the latter case – superstition or fear of "tempting fate" - was discussed in 2.2.1. Stuart et al. (2017) distance their experimental study from previous research, offering mixed support for the impact of desirability bias on perceived event probability (a separate mechanism), by theorizing that the tendency to prepare or not to prepare in anticipation of an event may be affected by manipulations of event desirability, independently from the impact on the perceived event likelihood. Hence, they distinguish preparation behavior as a stand-alone manifestation of desirability effects.

Lench (2009) also offers two possible explanations underlying how affective reactions may in practice lead to desirability bias in probability judgements (the idea of affective reactions was introduced in 2.2.2). First, (incidental and transient) affective reactions may have a general effect on choices, judgements and behavior, even when the emotionally-charged stimulus is unrelated to the event whose likelihood is being evaluated (her experimental study, see subsection 2.2.4, lends further support to this explanation). Second, and assuming affective reactions hold an even bigger sway on our judgements and behaviors, another possibility is anticipation as the underlying channel of effect of desirability bias. According to this interpretation, individuals anticipate their own potentially positive or negative affective reactions associated with the realization of the uncertain event and this leads to either approach or avoidance behavior. This approach/avoidance behavior towards an uncertain event is directly linked to its perceived likelihood estimation. For example, if an uncertain event into the future is anticipated to cause negative affective reactions to an individual (e.g. distress, sadness), then this individual would be psychologically predisposed towards avoiding this experience by judging it (Delusionally?) to be very unlikely to occur in the first place.

2.2.4. EMPIRICAL EVIDENCE

Bar-Hillel and Budescu (1995) identify three major types of events which are most often studied for the existence and intensity of desirability effects. These are personal events, social events, and aleatory (not to be confused with the economic interpretation; read further) or inherently neutral events. The first type, they elaborate, possess an intrinsic value of unequivocal emotional valence such as earning a promotion or contracting a deadly disease. The second type consists of events which are not universally recognized as either positive or negative - they are conferred a value and valence only through the means of subjective preferences e.g. election outcomes or competitive sports. The US 1932 presidential election was the experimental setting which Hayes (1936) used to demonstrate empirically for the first time that preferences and expectations are linked. An example of the final type can be drawing a specific card from a deck or drawing a marble of specific color from an urn. Research psychologist tend to expand on the interpretation of aleatory events which economists use and also characterize them with a neutral inherent desirability.

The three types of events also tend to be associated with different research paradigms, the authors observe. The first type tends to involve judgement of relative likelihood (probability that some event will affect the participant compared to another individual), the second often concerns predictions or expectations (predictions of emotionally invested participants are compared to indifferent participants), while the third often deals with events with known objective probabilities and seeks to compare how the perceived probabilities compare to those.

Windschitl et al. (2010) categorize relevant research on the existence and magnitude of desirability into four distinct research paradigms. Consistent with the established norm, desirability effects are contrived to the set of events which participants cannot expect to be able to influence. The four general paradigms concern either events which are stochastic (e.g. card-draw, see seminal study of Marks, 1951) or non-stochastic (e.g. competitive sports, see Massey, Simmons and Armor, 2012), while the dependent variable was either a likelihood judgement or a discrete resolution prediction. This approach reveals several insightful generalizations. First, by far the leading research paradigm in terms of the number of academic studies conducted involves stochastic events and discrete resolution predictions. Such studies (think of marked-card paradigm in the spirit of Marks, 1951) also produce the bulk of the significant effects. In contrast, the second most researched paradigm, involving once more stochastic events but likelihood judgement rather than discrete predictions on these events, rarely produces significant effects and usually of very small magnitude. An effect of slightly larger magnitude is generally discovered in the non-stochastic-likelihood judgement paradigm, whereas no studies were discovered that examine desirability bias in discrete resolution judgement tasks with non-stochastic events.

Hence, the strongest evidence to date regarding the existence of desirability bias comes from empirical studies in the spirit of Marks (1951) and mainly comes from experimental settings concerning decision-making under risk and not uncertainty. Windschitl et al. (2010) conclude that the findings from marked-card paradigm research experiments constitute the most convincing evidence in favor of desirability bias in forecast tasks. These studies involve dichotomous predictions about drawing a marked card from a deck of cards, usually with half the cards being marked, and show that participants tend to judge that the probability of drawing a marked card is higher than half when this event is associated with a monetary gain and less than half if this event is associated with a monetary loss (see Price 2000 p.162 for all other major studies replicating the seminal study of Marks).

Krizan and Windschitl (2007) conclude that despite the obvious appeal of the idea that desires bias optimism, the evidence supporting this is still limited (i.e. thin literature on the causal link between preference and expectation). As elaborated in Lench (2009), despite the abundance of correlational research studies, some important drawbacks in the methodology make it difficult to argue persuasively about the existence of a causal link between event desirability and bias in judgement or choice. Lench believes that one such drawback is the lack of distinction between the influences of the bias itself and the perceived likelihood associated with the event, e.g. severely undesirable events may be judged as less likely than moderately undesirable events simply because the former are objectively less likely to occur, and not because they provoke stronger bias in the response. Stuart et al. (2017) also find mixed evidence on the desirability-likelihood judgement link (significant but economically very small effects) and refer to studies questioning the existence of desirability bias in the context of expectations. However, they do find consistent evidence that people tend to show greater tendency to act in anticipation of desired events and lesser tendency to act in anticipation of undesired events.

Lench (2009) points out that most often in empirical studies the concept of desirability bias is associated with ex-ante intuitively predictable preferences over one event compared to another. A typical desirable outcome is often winning a game of chance, hence rendering trivial the interpretation of “desirability”. Besides noting the lack of more careful contemplation on what constitutes desirability, she also observes that these studies typically do not attempt to measure the degree to which participants desire or not the events studied.

In contrast, Bar-Hillel and Budescu (1995) argue in favour of the usefulness of such aleatory events in answering specific research questions about desirability bias. In the study of subjective probabilities, they argue that abstract, hypothetical or neutral events are popular based on the assumption that they do not affect utility and because often event desirability interacts with probability judgement. Hence, objectively equi-probable events with desirability or undesirability rating of identical magnitudes should possess identical perceived likelihood. However, pre-existing attitudes or convictions can confound the desirability effect. They study both aleatory and realistic events and manipulate desirability by either monetary or affective stimuli. No desirability effects are found for the former, while for the latter personal relevance (standing to gain) is found to play a larger role than simply attempting to evoke affect by manipulating the attractiveness of different events.

On the issue of confounds, Krizan and Windschitl (2007) summarize the empirical evidence on desirability bias and conclude that it is very pronounced for discrete outcome predictions, subjective probabilities do not seem sensitive to the desirability of chance outcomes (chance outcomes do not involve ambiguity), while when it comes to the desirability bias effects on bets the authors recommend caution. They argue that the two possible ways to make money with bets (one is when a prediction turns out to be correct while the other is when the desirable outcome occurs) introduces further confounding effects such as intentional, “normatively appropriate”, betting strategies (betting on the desirable outcome involves the potential of double reward) which may complicate the disentangling of desirability bias. It may also introduce other confounds such as superstition, e.g. betting against oneself.

On another note, Lench (2009) points out that despite the unquestionable rigor of research studies, such as Marks (1951), showing evidence of desirability bias in the context of chance events, they have been criticized for the lack of realism of their experimental setting. The argument goes that often participants are informed that each event is equally likely (hence, evidence is limited to 50/50 chance games setting) and the stakes are simply not that high compared to many situations which can occur in practice. It must be noted though that evidence on the effect of stakes on behavior and decisions is equivocal. Massey, Simmons and Armor (2011) similarly caution against readily generalizing past findings due to the unrealistically inconsequential nature of the desirability manipulations compared to what is often experienced in practice. Bar-Hillel and Budescu (1995) agree with the possibility that the desirability effect cannot be manipulated artificially, and that strong enough value manipulation cannot be achieved in an exogenous way (e.g. deciding arbitrarily which event is desirable by allocating a monetary reward to it cannot achieve a desirability effect of the needed intensity). Moreover, Price (2000) further question the Marks paradigm by pointing out that the little information provided to participants on which to base their judgements may have primed them into or gave them no choice but to behave in a way consistent with their wishes due to the lack of alternative approaches to follow. A further criticism of marked-card paradigm research experiments in the spirit of Marks (1951) is raised in Windschitl et al. (2010). They warn that an experimental design in which the participants are informed verbally about the value of marked and unmarked cards and subsequently have their forecasts elicited orally may be susceptible to experimenter bias and demand effects.

In contrast, and studying social events instead, Massey, Simmons, and Armor (2012) attempt to capture the relationship between very intense or strong desire and decision-making bias associated with football games predictions made by NFL fans. Controlling for team strength (confounding effect if stronger teams who are more likely to win are also more likely to be the favorite team of a given participant), familiarity (a famous results in this field is that people prefer betting on ambiguous events they feel competent in) and unbiased expectations (bookmaker predictions), they show that participants were ~16% more likely to predict a given team would win if that team was their favorite. The desirability bias hypothesis was upheld even when the favorite team regressor was replaced with a liking rating independent variable – a one division increase on the liking rating is associated with ~4% increase of likelihood to predict a given team would win. Similarly, they also show that a one division increase on the desirability response form (measure of how strongly each fan wanted their favorite team to win the next game) was associated with 5% greater likelihood to predict a win for the favorite team. By interacting the predicted point spread (as an indication of the degree of ambiguity) with the favorite team regressor, they show that desirability bias exists along the entire unbiased distribution of likelihood but is most pronounced around 50% (greatest ambiguity). They raise a question mark over their findings concerning the significance of the provided accuracy incentives, and hence whether participants indeed believed in their predictions (loyalty vs hedge against pain).

In the same spirit as Massey, Simmons and Armor, Price (2000) sets about testing the existence of wishful thinking by achieving a more effective manipulation of participants' desires. His experiment involves a realistic setting in which participants are randomly allocated to one of two groups competing in a dart-throwing contest. He argues that the desire associated with the potential victory of the group one is affiliated with is stronger or at least of a different kind to that elicited by other laboratory experiments that instead endow arbitrarily one event with a monetary gain. A mean rating across the two teams of 1.8 on a 0-3 rating format measures the desirability that a team member would hit the bullseye. The statistical analysis shows that the desire is significantly associated with the predicted score even after controlling for team affiliations, suggesting that the team effect alone cannot explain the judgments.

More on social events, a study by Bar-Hillel, Budescu, and Amar (2008) also investigates the existence of desirability bias in the field of competitive sports since there "wishful thinking seems rampant" (p.279, 2008). They similarly control for confounds such as knowledge and pre-existing preferences (exclude certain data points based on whether the experimental desirability manipulation coincides with the existence of prior non-neutral attitudes). Moreover, they also test an alternative explanation for their results – that of salience or pure attentional factor. In other words, they consider if the act of drawing attention to the event selected at random to bear monetary rewards may be enough elicit desirability effect. It turns out their statistical analysis shows no significant difference between the distribution of the probability estimates provided in the two conditions, i.e. desirability and salience manipulations. Hence, they reject the motivational mechanism of desirability bias due to the identical magnitude of the effects in both treatments and the inability to manipulate event desirability while holding event salience constant.

Such results are consistent with what they argued in the past. For example, in Bar-Hillel and Budescu (1995) they zoom in on different studies examining biased attention to evidence as a possible cause of wishful thinking in response to a possible limitation that the desirability effect is weaker when all relevant evidence is presented in front of the participants. Hence, they dispute previous findings in the literature where participants are allowed to retrieve and attend to evidence at will and claim that what these studies find is not biased probability judgement but rather an unbiased probability evaluation based on a biased body of evidence. On that note, Fischer and Budescu (1995) support this notion in their interpretation of desirability-prediction link in the context of Israeli voters, arguing it is selective social interaction which promotes distorted likelihood evaluations rather than the desirability effect (see also Uhlaner and Grofman, 1986)

Finally, Windschitl et al. (2010) conduct a series of experiments aiming to test the psychological drivers behind the desirability effect as well as explicitly investigate their implications for the existence and magnitude of desirability effects in different experimental research paradigms (not just marked-card). They tested experimentally the predictions of four desirability bias accounts in research paradigms involving non-stochastic events and likelihood judgement tasks in addition to the classic stochastic event-dichotomous prediction task paradigm. The authors replicate the desirability bias in the marked-card paradigm even after correcting for some potential pitfalls in previous studies which may have led to experimenter bias and demand characteristic effects, hence providing evidence against the artifactual account of desirability bias discussed in 2.2.3. In their study, the value of a marked card is not known by experimenters and, instead of having the experimenters elicit the predictions orally, forecasts are entered into a computer terminal. In experiments two and three, they test for the existence and magnitude of desirability bias in the non-stochastic event-dichotomous prediction task paradigm and use the findings to determine the dominant underlying psychological mechanism. The crucial piece of evidence from these experiments is the reappearance of desirability bias with non-stochastic events under perceived balance of evidence (no desirability effects when participants did not feel it necessary to make a subjectively arbitrary forecast). Finally, in their fourth experiment, Windschitl et al. confirm the presence of desirability bias in post-diction tasks, similar to prediction tasks, and register a substantial drop in magnitude of the effect when subjective likelihood judgements are elicited as compared to dichotomous predictions. Interestingly, the authors show that the desirability bias reappears even in the subjective likelihood judgment task if subjects are encouraged to rely on their “hunch” or guess.

The authors argue that the results from their experiments lead to the conclusion that people may not exhibit large-scale desirability bias and instead be influenced by their desires only in situations encouraging them to rely on their hunch or gut feeling. The experiments of Windschitl and collaborators provide strong evidence that the main psychological driver behind the desirability bias found in marked-card paradigm experimental studies is indeed biased guessing. This effect holds more broadly for stochastic events in both prediction and post-diction tasks. Biased guessing is found to be less relevant for non-stochastic events since there pre-existing level of perceived competence, and hence biased evaluation of evidence, plays a much bigger role. However, their experiments showed that prediction can be susceptible to subjectively arbitrary guessing, and hence biased guessing, even with decisions concerning non-stochastic events, so long as these events appear virtually identically likely to the participant (exceptionally difficult trivia questions serve this purpose in their study).

3. EXPERIMENT: DESIGN AND METHODOLOGY

Section 3 is organized as follows – 3.1 introduces the matching probabilities method developed by Baillon et al (2017) to measure ambiguity attitudes, 3.2 explains how the concept "event (un)desirability" is operationalized in this study, 3.3 presents the experimental groups or conditions, while 3.4 deals with further, less-central though still important, experimental design choices. Subsection 3.5 attempts to piece together several tentative hypotheses based on the theory and empirics reviewed thus far.

Readers are free to skip the parts detailing background information behind the methodology and experimental choices. These excerpts are preceded by the text "*Background information*" and are placed at the beginning of each subsection, where applicable. Readers are then suggested to proceed towards the end of the relevant subsection where the text "*Application to the present study*" marks the beginning of explanations detailing the specific methodologies and features applied in the present study.

3.1. MEASURING AMBIGUITY ATTITUDES USING THE MATCHING PROBABILITIES METHOD

Background information:

In this study, ambiguity attitudes are operationalized following the method proposed by Baillon et al. (2017). They propose unifying indexes used by researchers in the past into two dimensions - aversion and insensitivity. This is argued to offer a more complete explanation of attitudes towards uncertain events. Aversion in this context has an identical interpretation to the one used in the domain of choice under risk while insensitivity captures the perceived degree of uncertainty or how well the subject can discriminate between different levels of probability. Baillon et al. cite empirical findings supporting the notion that the first index (aversion) is dependent on the second (insensitivity) which leads them to conclude that failing to account for likelihood can produce misleading predictions about ambiguity aversion.

Baillon and his collaborators postulate that their method requires six questions (plus an extra two if the researcher wishes to employ consistency checks) in which the subjects are asked to choose between (bet on) a risky and (or) an uncertain event. Additionally, the events chosen must be of moderate likelihood. They define their method in terms of gains, however it can similarly be applied to the domain of losses. All events considered must come from the same source of uncertainty and be mutually exclusive and exhaustive, i.e. the realization of one scenario must preclude the realization of other scenarios and all events together describe the whole set of possibilities. This is the key assumption of their method that allows for the distinction between different attitudes towards ambiguity and their isolation from the perceived levels of event likelihood. In this context, ambiguity aversion is defined as:

$$b = 1 - \overline{m_c} - \overline{m_s}$$

where the middle term on the right is the average of the matching probabilities of the three single events (i.e. $\overline{m_s} = (m_1 + m_2 + m_3)/3$, where $m_i = m(E_i)$), while the rightmost-term is the equivalent for all composite events (i.e. $\overline{m_c} = (m_{23} + m_{13} + m_{12})/3$, where $m_{ij} = m(E_{ij})$). As the authors elaborate, if a subject is ambiguity-neutral then, by definition, the subjective probability of an event is identical to its matching probability, i.e. $m_i = P(E_i)$ and $m_{ij} = P(E_i) + P(E_j)$. The matching probabilities of the uncertain events are elicited in this manner for all single and composite events. The equation above implies upper and lower bounds for b equal to 1 and -1, implying maximum ambiguity aversion and maximum ambiguity seeking, respectively. To quantify the perceived distinction between different levels of likelihood the authors propose examining the (conveniently rescaled) difference between average matching probabilities of the single and composite events:

$$a = 3 * \left(1/3 - (\overline{m_c} - \overline{m_s})\right)$$

This rescaling yields upper and lower bounds of 1 and 0, implying complete insensitivity and discrimination respectively. This is so because ambiguity neutrality and perfect discrimination yields $\overline{m_c} = 2/3$ and $\overline{m_s} = 1/3$. As a side note, recall from subsection 2.2.4 that the presence and magnitude of event desirability effects may depend on the choice setting of a task. Therefore, one may speculate at this point that the matching probabilities method for eliciting ambiguity attitudes may introduce complications (i.e. lead to weaker results) because it involves asking subjects to place bets. In contrast and based on what we know from past empirical findings discussed in 2.2.4, an alternative method asking subjects to make subjective likelihood judgements may establish stronger results. In the context of this study subjects are asked to bet on future events rather than on knowledge of past events.

The advantage of this method is that one can always separate respondents' ambiguity attitudes from their subjective probability evaluations because a-neutrality is calibrated, i.e. the values 0 and 1 have consistent meaning irrespective of likelihood judgments. This then enables the researchers to use natural events for measuring ambiguity attitudes – a much needed development given the existing gap in the literature. Moreover, Baillon et al. also argue that their method is valid for a number of different theories, and hence does not require that behavior abides by the rules of expected utility. Finally, by using the matching probabilities method to estimate their ambiguity indexes, the authors argue that they can effectively isolate ambiguity from risk preferences (see Dimmock, Kouwenberg, Wakker, 2016).

Application to the present study:

Following the method as explained earlier in this subsection, from each source of uncertainty examined in the present study I extract three single events which then also form the basis for the construction of further three composite events. Careful attention was given to splitting the source in a way that the three events chosen all appear of moderate likelihood and approximately equally likely. This is a typical characteristic of aleatory events as interpreted and studied by economists but can also apply to cases in which the sources where events are extracted from are associated with intensive intrinsic meaning (e.g. social events as referred to in the psychological literature) provided the weight of evidence favoring or disfavoring each event is balanced (i.e. eliminating the influence of perceived knowledge). You may recall from 2.2.3 that this situation facilitates one of the most popular mechanisms through which desirability bias may operate – resorting to subjectively arbitrary choices which often lead to erring in the direction of what appears to be a desirable resolution. However, the design of this study explicitly measures the effects of perceived competence on the topic at hand (this issue is explored in the next subsection – 3.2) while the matching probability method controls for the effects of subjective likelihood evaluation, and hence the present research is insured against such concerns.

With current probability answer options, the a-aversion range is from 0.85 to (-0.85) while calibration still occurs at 0. This does not necessarily imply ambiguity neutrality though, e.g. single and composite event matching probabilities of 0.5 would give an a-aversion index of 0 and a-insensitivity index of 1 while ambiguity neutrality also requires that $m_i = P(E_i)$ and $m_{ij} = P(E_i) + P\left(E_j\right)$, which then automatically leads to $b = 0$. This restricted interval is due to the precision level of the matching probabilities, i.e. smallest interval is 10% and no follow-up questions are asked to pinpoint the matching probability with greater accuracy. This creates the peculiarity that if a respondent always prefers the risky bet over the uncertain bet for either single or composite events, then the corresponding matching probabilities would be 5% and 10% respectively (i.e. the interval midpoints), resulting in a-aversion index of 0.85

In a similar fashion, a-insensitivity is instead calibrated at 0.3 instead of at (1/3) because the level of precision of the matching probability intervals implies that the a-neutral single event probability is 0.35 while for composite events it is 0.65, while it is standardized to 1 by multiplying it by (10/3) rather than 3 (so that a consistent meaning is again attached to the values of 0 and 1). Then, a score of 0 indicates perfect discrimination (but not necessarily ambiguity neutrality, e.g. a respondent with sing le and composite event matching probabilities of 0.45 and 0.75 would also have an a-insensitivity index of 0 but would be considered ambiguity-seeking and have a negative a-aversion index) while a score of 1 points to a complete lack of discrimination There are also instances of the index going above 1 – in those cases respondents have provided answers which result in the average probability score for the single event to be greater than that of the composite event, i.e. $\overline{m_c} < \overline{m_s}$. In other cases, the index goes below 0 – the area beyond perfect discrimination and of underestimation of the probability of single and overestimation of the probability of composite events, i.e. $\overline{m_s} < 1/3 < 2/3 < \overline{m_c}$. Nevertheless, all such responses were kept in the sample. Only respondents who switched multiple times their choices between the ambiguous event and the risky event for different degrees of probability were excluded from the final sample.

3.2. OPERATIONALIZING THE EVENT DESIRABILITY CONCEPT

Background information:

As mentioned in section 2.2, there is no universally agreed upon definition in the literature of what constitutes event desirability, and hence different methods of operationalizing this concept have been proposed so far (Lench, 2009). However, at the most general level, two such approaches to measuring desirability effects can be identified. Those can be classified as either making use of “intrinsic” (usually nonmonetary) or “extrinsic” (monetary) desirability design. What immediately distinguishes these two approaches from one another is the type of events used in the experiments. As discussed in section 2.2.4, there are three general types of events studied in this literature – personal, social, and aleatory. The former two have an inherent and subjectively perceived degree of desirability attached to them – i.e. intrinsic – whereas the latter one typically involves conferring meaning to an otherwise neutral event.

Additionally, recall from section 2.2, an alternative way of classifying the experimental research done so far is offered in Lench (2009). She distinguishes two conflicting schools of thought that treat desirability either as a phenomenon which is constant between subjects or as a dynamic phenomenon influenced by unpredictable and transient affective states. Arguably, Lench implies that the former conceptual treatment ascribes to a given event a consistent emotional valence (not necessarily true for magnitude however) across individuals. An example of such an experimental design is Windschitl et al. (2010), who manipulate desirability of different ambiguous event resolutions through monetary means, i.e. the decision-maker receives a monetary amount if an arbitrarily chosen (aleatory) resolution occurs, irrespective of their choices.

However, one should not explicitly associate this section of the experimental literature, i.e. studying aleatory events, with the treatment of desirability as a constant phenomenon. As an example of a similar experimental design, which is however closer to the conjecture that event desirability is of ephemeral nature, is the study of Lench (2009) herself. She conducts an experimental study to manipulate the desirability of stimuli by pairing an initially neutral-desirability event with either positive or negative stimuli to create affective reactions of varying valence. In her example, participants were asked to judge the likelihood of owning a white car in the future (neutral-desirability event) and the likelihood of experiencing minor positive and negative events compared to their peers. Immediate affective reactions, supposed to influence the degree of desire associated with the neutral event, were elicited beforehand by pairing the image of a white car with pictures of either inherently positive or negative stimuli (desire created by approach/avoidance for situations which in turn are evoked by positive/negative affections). This design also arguably falls under the category of studies on extrinsic desirability because the non-neutral stimuli have expected and predictable affective responses (constant/universal “value”) between subjects (similarly to positive monetary rewards). However, it also demonstrates that different affective states can confer the same stimulus varying degrees of desirability.

In such studies, the experimenters do not usually need to explicitly measure the valence and intensity or magnitude of the stimulus since there should be a relatively small window for discrepancies in evaluation (for example in the perception of monetary rewards) while the aleatory events used, by definition, carry no prior emotional baggage. In contrast, this would not be so for studies taking the intrinsic desirability route. For example, aleatory events would not be a suitable choice for an experimental design making use of intrinsic event desirability. Nevertheless, they are a common way of operationalizing event desirability because prior attitudes are mitigated. Hence, one can avoid the more complicated procedure of attempting to measure these prior preferences (i.e. ranking and rating of the desirability of ambiguous events). In contrast, studying event desirability with personal and social events suffers from this design peculiarity. An example of such a study and how this issue is commonly solved is Bar-Hillel, Budescu, and Amar (2008). They control for pre-existing preferences in their desirability manipulation of competitive sports (social event) predictions by excluding from the study participants who report loyalty to a team subject to desirability manipulation. In this design, a monetary reward (the *de facto* desirability manipulation) of 4-5$ is attached at random to an event where a given team wins their game, while the self-reported fans of that team are excluded from the study to avoid potential confounding effects. A second payment of equivalent value is promised to the participant correctly predicting the greatest number of games, representing their precision incentive to encourage optimal effort and effectively creating conflicting incentives.

However, their study would have looked differently if they had followed an alternative approach in which the self-reported sub-sample was included in the final sample of subjects. Arguably, they take the extrinsic desirability route in their experiment instead of taking full advantage of the social event they study. Tappin and van der Leer (2017) is an example of this alternative approach. They also study social events but employ intrinsic event desirability methods by creating a 2x2 quasi-experimental desire-belief groups design from a sample of Trump and Clinton supporters in the 2016 US presidential election. They achieve this by sampling their subjects’ political affiliations and subjective probability evaluations of each candidate’s chances of becoming the new US president prior to administering the treatment. The two distinct desirability groups (Trump and Clinton supporters) are randomly treated with confirming and disproving information conditions, so that approximately half of each of the two groups receive the confirming while the other half – the disproving treatment. They use this design to measure the relative magnitudes of desirability confirmation bias effects on the subjective probability judgements.

The present study resembles more the approach used by Tappin and van der Leer but departs slightly from it in a sense that when examining the relative magnitude of the perceived competence and desirability effects on ambiguity attitudes it does not block by the covariate (i.e. competence). Instead, this study samples the magnitude and direction of the hypothesized covariate, retaining statistical power and exogenous allocation to conditions. In contrast, the alternative design (i.e. the one used by Tappin and van der Leer) would have involved forming two groups of high-low competence (hence, sampling competence before allocating to groups) and administering the desirability treatment randomly to half of each group and the undesirability randomly to the other half. This creates a 2x2 design with 4 between-subjects conditions. Krizan and Windschitl (2007) believe that there are two main reasons why going down the intrinsic path to operationalizing desirability is useful. First, the number of possible alternative desirability manipulations is much larger within the nonmonetary compared to monetary domain. Second, the desirability effects caused by nonmonetary manipulations may be very different from those generated by monetary incentives. This conjecture is also tested in Price (2000), and more specifically the possibility that laboratory manipulations cannot reproduce the same intensity of desire often associated with real-world events (e.g. the success of the favorite presidential candidate or sports team). However, the main obstacle to such studies remains the potential for confounding factors (most often expectations) that are not controlled for (or uncontrollable).

Therefore, to claim the relevance of desirability bias using the intrinsic desirability approach (and studying personal or social events), other potential drivers of ambiguity attitudes should be controlled for. As another example, Heath and Tversky (1991) use sorting of subjects according to their expertise. In their procedure, subjects are initially asked to choose a topic they feel most strongly about (additionally they may also be asked rate the intensity of this feeling) and are then sorted between the different conditions according to their responses. Heath and Tversky then use only subjects whose competence self-ratings lie on the opposite ends of the midpoint between the two events studied. Another solution is to ask the subjects to rank the topics from most undesirable (from negative through neutral/no preference to positive) to most desirable, or simply choose the most desirable and most undesirable events. Such a design could easily be replicated to answer the research questions of this paper. However, it would also be guilty of administering the treatment in an endogenous manner. There are also further potential design flaws. For example, the topic ranked as most (un)desirable may still not carry significant emotional investment, and hence the need for feelings intensity follow-up question for both extreme ends. For example, any score above or below the midpoint on the feelings intensity response spectrum would then qualify the subject to the relevant condition. These are some of the most important considerations which influenced the final design of the present experiment.

Application to the present study:

Considering earlier findings, in this experimental study (which arguably also makes use of social events, i.e. such that receive an emotional valence label only through subjective evaluations) the competence and desirability effects receive all the attention as the main drivers of ambiguity attitudes. Arguably, the comparative ignorance hypothesis (CIH) and home bias would be irrelevant in this design as subjects are not encouraged to think about experts and the knowable but missing information, and not encouraged to compare knowledge of domestic to foreign events. Moreover, comparative ignorance in this context would be identical to competence because conceptually the former exists because of the latter (Fox and Tversky, 1995). The present study also departs from the covariate-blocking approach in that when examining the relative magnitude of the perceived competence and desirability effects on ambiguity attitudes it samples the magnitude and sign of competence but does not use their answers to allocate subjects to conditions, hence retaining statistical power and exogenous administering of treatment. Therefore, the study can distinguish the impact of desirability bias using the intrinsic desirability approach (and studying personal or social events) because the effects of other key potential drivers of ambiguity attitudes are measured.

Further, desirability (or undesirability) ratings were collected for all single events. I employ a relatively unsophisticated method for measuring (and interpretation of) event desirability. In the spirit of studies following the intrinsic desirability route, I use a simple 7-point Likert-type scoring response format with a numeric midpoint of zero and extending across the domains of negative and positive numbers. Respondents were gathered online and presented with one Likert-type response item per single event asking them to determine how strongly they wish or do not wish that the present event comes true in the future. The answer options ranged from -3 (strongly undesirable resolution or event) to +3 (strongly desirable resolution or event), with 0 being the numeric (and psychological, as I would later argue in subsection 3.4) midpoint. According to the cautionary words of Bar-Hillel and Budescu (1995), the elicitation procedure is of great importance as there is the danger of data contamination in the case of biased question framing. To this end, the subjects in this study are not asked how desirable something is as it may prime them to engage in selective search and evaluation of arguments favoring the view that the event is indeed desirable. Instead, a more neutral version of this statement is used to confer the message, encouraging them to think about how strongly they desire an event to occur or how strongly they desire the same event not to occur. Hence, I attempt to sample the subjective intrinsic (un)desirability of social events.

Generally, the competence and ambiguity aversion indexes for example are "global", in a sense that they relate to attitudes at the source of uncertainty level, whereas desirability bias is almost always operationalized as a "local" index - relating to attitudes towards specific resolutions or events contained within that source. To give an example using the present study, if a respondent indicates a given degree of competence in correctly judging the likelihood of the event in which less than 15% of the global energy mix in 2025 is comprised of renewable energy sources, then it is reasonable to assume that this respondent also possesses the same degree of competence in judging the likelihood that this share ends up anywhere between 15% and 100%. Hence, his or her degree of competence extends to all events within the given source of uncertainty – it is a global attribute and one question "at the source level" should suffice to measure it. This type of operationalizing the perceived competence concept is similar to what is found in Kilka and Weber (2001) for example.

Normatively speaking, this should also hold for the desirability measure, provided we study mutually exclusive and collectively exhaustive events since such events cover the whole spectrum of probability from zero to certainty. Hence, from this perspective, taking two events from the same source, it would appear irrational to evaluate as desirable the event described by a probability interval including zero probability and the complementary event (described by probability interval including certainty). In plain parlance, this would mean that the person wishes the event to occur and not to occur at the same time (*This would only be possible under the laws of quantum mechanics!*). Instead, the desirability evaluation of an event should go together with an identical undesirability evaluation of its complement. In other words, examined together, the desirability evaluations of the set of all possible resolutions and their complements for a given source of uncertainty should result in zero (global) expected score.

However, what if respondents are influenced by an "overarching desirability attitude" towards the source of uncertainty itself, and hence this biases their event desirability evaluations away from rationality (here defined as zero expected score)? Then it would not be possible to infer in a straightforward way and based simply on their evaluation of one possible resolution their desirability evaluation for its complement. Hence, one can speculate that if the examined events within a given source of uncertainty produce ratings on the opposite sides of the midpoint (inversely related) then this would provide a weaker case for the potential impact of desirability on ambiguity attitudes because it would be evidence for rationality as defined earlier.

To illustrate, imagine being asked whether a resented political dictator was born on an odd or even day – evaluations satisfying such a pattern would be a positive desirability score for one event (e.g. odd day) and a negative desirability score for the complement and opposite event (even day). Looking at what may happen to the matching probabilities of the events in such a case, and assuming that desirability reinforces ambiguity-seeking behaviors (which might well not be so as explained in numerous places in section 2), desirability bias for one event (e.g. odd day) pushes its matched probability up, whereas the undesirability bias for the polar opposite (even day) pushes its matched probability down – hence, the final impact of the bias is not immediately clear. As will be explained later (subsection 4.1.1) some take this argument further to justify using single-item, per-event scores as opposed to scales or global desirability evaluations. They attach little meaning to such summated desirability scores precisely because they conceal the underlying variability in responses (especially in cases where the expected score is zero, e.g. if one event scores a 3 on the undesirability-desirability spectrum while its complement scores a -3) and question what the potential insights gained from such total score or averaged scores.

In this study, I find it insightful to explore whether such overall negative or positive attitudes towards the source can exist, manifested in the sign-consistency of event desirability scores (hence, a non-zero expected evaluation), by selecting sources of uncertainty which should provoke (based on feedback from piloting phase) consistent-valence stimulus evaluations (i.e. encourage all-positive or all-negative desirability scores) for all events contained within. In other words, if a contrasting pattern in the evaluation of opposite events can only be seen between subjects but not "within subjects", then this may be suggestive of the presence of a global "source desirability" which may be influencing (or biasing) the event desirability ratings.

However, it may also be the case that the stimuli turn out to be bipolar, i.e. sources turn out to be of bipolar valence. Then this would be evidence against such a global influence and potentially against biasing. Such a hypothesis (i.e. the presence of rational and irrational part to desirability ratings) can also be confirmed or disproved empirically using factor analysis and other unidimensionality tests. This may prove a fruitful area for further research. One can speculate that such a systematic effect on event-specific ratings is an important component of desirability bias. It is of course also possible that the ambiguity attitudes may also be affected by the local desirability evaluations of single events, which is a conjecture that I will also be testing in section 5. Nevertheless, such speculations are of tangential importance for the purposes of the present paper as it is beyond the scope of this work to try and test alternative theories of the psychological origins or drivers of desirability bias. The main goal is therefore to investigate the relationship between ambiguity attitudes and event desirability.

3.3. EXPERIMENTAL GROUPS

The sources of uncertainty studied in the present experimental undertaking contain, by definition, social events. However, these events arguably have a universally agreed upon emotional valence that the approach to operationalizing event desirability may in practice be found equivalent to the extrinsic. In this sense, this study is similar to Lench (2009) in that the stimuli used should have predictable affective responses. This allows for an exogenous, random allocation to the control and treatment conditions. I assume that the stimuli should have a consistent evaluation valence between subjects which makes redundant the measurement of prior predisposition to these stimuli (beyond regular “due diligence” in the form of piloting, of course).

Therefore, and in contrast to the study by Bar-Hillel, Budescu, and Amar (2008), I do not attempt to eliminate pre-existing preferences but rather use the reported intensity of this preference as the main independent variable in my analysis. I expect that responses from subject who are randomly allocated to the control condition would have a frequency distribution concentrated around a neutral stimulus evaluation (the 0 on the undesirability-desirability Likert scoring response format), subjects who are allocated to the desirability condition would provide responses distributed in the positive end of the choice spectrum, while the rest of the respondents should give a predominantly negative evaluation. Therefore, the random allocation to either the control or any of the two treatment conditions represents the desirability manipulation. Subsection 3.1 explained the method used in this study which enables this experimental design and the study of ambiguity attitudes with natural events. Recall from section 2.1 that a defining feature of the experimental literature on ambiguity attitudes is the study of artificial events because of the difficulty of distinguishing true ambiguity attitudes from differences in the perception of likelihoods in the case of natural events.

The source of uncertainty examined in the neutral or control condition (labeled “condition 0” in statistical output) contains events related to different future scenarios regarding the comparative height of an under-construction skyscraper in Jeddah, Saudi Arabia (disclaimer: experiment completed before the Khashoggi incident). The source of uncertainty examined in the desirability treatment condition (carrying the numerical label “2” in most statistical tests I perform in section 5) contains events related to different future scenarios regarding the share of renewable sources in the global energy mix. The source of uncertainty examined in the undesirability treatment condition (with numerical label “1”) contains events related to different future scenarios regarding the share of jobs globally becoming automated (i.e. the process of labor displacement by artificial intelligence and robotics). These sources were selected after extensive piloting (i.e. approaching potential respondents at random and performing cognitive interviewing on the meaning of the items in the questionnaire) among students at the Erasmus University campus. The extended version of the questionnaire (with all branches and sections shown) is available in the appendix.

Each respondent evaluated the Ellsberg’s urn problem and one randomly assigned natural event depending on which experimental group he or she was allocated to. Although the administering of the treatment is on an arbitrary basis it is still ensured that respondents are split approximately evenly across the three conditions. Additionally, the order in which the ambiguity indexes elicitation and competence or desirability measurement are presented is randomized. Not randomizing could lead to behavior according to stated competence or could provoke anchoring in event evaluation. During the piloting stage I discovered that respondents seem tempted to always give a maximum desirability rating to the events they already placed a bet on, irrespective of how desirable they find the event in general. Hence, placing the competence and desirability rating after, rather than before matching probabilities have been elicited, does not seem like an innocuous ordering. In any case, and in addition to the order randomization, I include a sentence to inform participants that it is important to ignore any bets they may have placed before. The competence and desirability measurements are also presented in random order in order to mitigate potential bias favoring reasoning or gut feeling strategies.

3.4. FURTHER EXPERIMENTAL DESIGN FEATURES

This subsection covers issues such as suitable design choices in Likert-type questionnaires to ensure quality data, the use of question order randomization, and the process for of acquiring participants, among others.

When it concerns Likert-type response formats (the ones used in the present study), Rossiter (2002) levies heavy critique on placing the intensity of a Likert-type item into the “stem” (i.e. question itself) as opposed to into the “leaves” – answer options. He argues that this method causes the distinct answer options to have an ambiguous, and hence inconsistent, interpretation across respondents, prohibits the existence of a true and valid psychological zero in the spectrum of answer options, and obscures the meaning of any form of aggregated total score. For example, he asks, what does “Strongly disagree” mean when the item stem presents the statement “I never attend social events” compared to “I seldom attend social events.” (p 34)? Does the respondent disagree because she always attends them or because she attends at least a few? The psychological zero is also lost in both versions of the item stem because “neither agree nor disagree” has different meaning in each of them. (Intensity-)Neutral item stems are employed in the questionnaire used to gather data for the present study. This, in addition to building a spectrum of categories ranging from minimum to maximum intensity, is the recommendation of Rossiter. However, it could also be argued that there is no amount of objectivity and clarity that can be built into the stem or leaves in Likert-type response formats to eliminate any reasonable doubt of heterogeneity in the interpretation and perception of items across respondents.

According to Rossiter, there are three main types of answer option formats which concern the characteristic of the object being studied. These are probability, frequency, and degree. It is also generally accepted, as noted by Rossiter, that, for numerically-labeled answer options, five (for unipolar) to seven (for bipolar spectrum, ideally extending into negative and positive number domains) distinct categories reach the upper limits of reliability and psychological discrimination (Allen and Seaman, 2007). This implies that adding extra answer options increases discriminatory power less than it increases the cognitive and time demands on respondents (to adequately process them). This tradition has held out remarkably well since it was first popularized in Miller (1956). All features mentioned so far are part of the questionnaire used for the present research.

Further, the expected worst-case event is always presented first when desirability is elicited in order to try and encourage absolute scores and not relative. It is important, I speculate, that subjects feel encouraged to consider the desirability of each event independently in order to provide scores which reflect more closely the intensity of their feelings towards the stimulus. In contrast, providing relative desirability scores obscures this information. In this context, it was reasoned that an experimental design which presents the "worst" (or "least best") event always first would prevent inflated scores of respondents who evaluate the scenarios relatively. If instead they were presented with the most favorable scenarios first, then the least favorable would tend to receive an exaggeratedly negative score. However, another approach could have been to always present first the one event predicted to provoke the strongest reactions, i.e. worst event in undesirability and best event in desirability. The question which approach would achieve the desired results is inevitably answered mainly through trial-and-error.

I also opted to exclude attention checks from the experimental design due to the further complications that they can introduce (e.g. some claim that excluding respondents on the basis of failed attention checks can lead to bias in the data - Berinsky, Margolis, and Sances, 2014). It is agreed that passing attention checks should not necessarily be expected to signify greater effort or non-random response patterns.

Furthermore, subjects were recruited online via social media websites (mostly through Facebook), messaging platforms (e.g. WhatsApp) and survey sharing platforms (SurveyTandem). Subjects were not compensated (i.e. not provided with a monetary reward) for participating or based on performance. However, this should not necessarily imply that subjects exerted sub-optimal effort as all of them participated on a voluntary basis (e.g. studies have shown that no pay is better than low pay - see Gneezy and Rustichini, 2000). Aleatory events (i.e. not the ones studied in the present paper) are the ones that are generally associated with incentives for good performance in the psychological experimental literature.

Additional data was also gathered such as respondents' gender, age, degree, and the time they took to complete the questionnaire. Incomplete answers were not used in the analysis. Arbitrary answers were also not retained. The latter could easily be recognized as they all followed a similar choice pattern on the matching probabilities part of the questionnaire – preferences between the risky and uncertain event switched back and forth forming a zig-zag-like shape. It is evident why such a response pattern constitutes low quality data and is thus eliminated from the analysis. The total sample size which ended up being used is comprised of 102 responses.

3.5. HYPOTHESES

Although this research should perhaps best be categorized as exploratory in nature, I nevertheless tentatively form several expectations based on the theory and empirical evidence reviewed so far. These relate to how event (un)desirability evaluations will be distributed across the three different experimental groups (studying neutral, desirable and undesirable stimuli) and how these conditions will differ in terms of their ambiguity attitudes (as measured by two ambiguity indexes).

First, in the control group event (un)desirability evaluations should cluster around the midpoint zero of the answer spectrum, i.e. the neutral desirability answer options should be the most chosen one. The undesirability condition should approximate a positively skewed distribution in which negative evaluations outnumber positive, while the opposite of that should be observed in the desirability treatment. These three broad propositions are henceforth referred to as the “secondary” hypotheses because they are only peripheral to the main research questions.

Second, holding perceived likelihood and competence constant (as well as other relevant controls), I hypothesize a lower ambiguity aversion index for desirable events and higher ambiguity aversion index for undesirable events, compared to a control group without desirability manipulation (or neutral desirability manipulation, depending on the perspective). So, even when an ambiguous event is perceived to be as likely as a risky event subjects still prefer to bet on the former to a greater degree if it is especially desirable and vice-versa. In other words, desirability bias has an impact on ambiguity attitudes beyond mere likelihood distortion effects and may have the power to outweigh the competence effect (in cases when they work in opposite directions). On the other hand, Stuart et al. (2017), for example, posit that superstitious beliefs such as fear of tempting fate (Risen and Gilovich, 2008) may cause pessimistic choices in the context of desirable events. For the present study this would mean that we should observe a higher a-aversion index for desirable events and lower a-aversion index for undesirable events compared to the control group. However, evidence for such attitudes is scarce in the literature, and therefore I judge the former cases to be more likely.

Third, a-insensitivity may also be affected by a different factor, for example such as the dispersion of the desirability scores (within-subject and -source, between-events), rather than any measure of central tendency as with the a-aversion. It is perhaps reasonable to assume that the less dispersed the desirability scores of a respondent are (measured by nonparametric statistics such as the absolute deviation from the median between-subject rating or the range of the within-source ratings) could lead to lesser degrees of discrimination between single and composite events.

The second and third broad hypotheses concern the impact of event (un)desirability on the ambiguity attitudes and are therefore labeled “primary” research questions of interest.

4. ANALYSIS OF THE DATA

The split into “background” and “application” text remains relevant for this section.

4.1. EXAMINING THE NATURE OF THE DATA (REVIEW OF LITERATURE)

The statistical methods used here broadly serve two purposes and form a two-step process of data analysis consistent with the split of hypotheses into primary and secondary. First, determining whether the observed scores for the experimental independent variable measuring event (un)desirability indeed follow the hypothesized directions (subsection 4.2). Second, whether the observed indexes measuring ambiguity attitudes behave in a way suggesting that the treatment condition indeed provokes a change in behavior compared to the control (and Ellsberg) condition (subsection 4.3). And finally, though not the focus of this paper, I will also spare a few thoughts on whether the results provide a hint of evidence favouring or disfavouring any of the theoretical frameworks reviewed earlier.

Before that, I begin by exploring the issue of choice between single- and multiple-item response formats to measure the main independent variables (4.1.1) and explain the context behind the choice of specific tests to used obtain evidence regarding the main hypotheses (4.1.2).

4.1.1. ON THE APPROPRIATENESS OF SINGLE- VERSUS MULTIPLE-ITEM LIKERT-TYPE MEASURES

Background information:

As mentioned in subsection 3.2, this experimental study uses single-item Likert-type responses to measure the main independent variables – event (un)desirability and perceived competence. This may also be the right place to define Likert Scales and the distinction from individual Likert-type items (method followed in this paper). In the seminal work of Rensis Likert (1932), he implies that *scales* mean the aggregation of several (Likert-type) items or questions measuring the same underlying variable or construct. By the original definition, a single-item measure is never a scale, and therefore an alternative term should be used instead such as a Likert-type item (or response format). Aggregating should in theory eliminate idiosyncratic error variance and produce data approaching interval measurement scale. More on the type of data produced by Likert-type response formats follows in subsection 4.1.2. When several items are theorized to measure a single construct, then usually aggregating is done via means of structural equation modelling (SEM) as it is not best practice to simply perform aggregating arithmetic operations on the data (e.g. mean, sum). Depending on the assumption made about the underlying structural relationship of the data and the aims of the researcher, the two most established groups of models are principal component analysis (PCA) and factor analysis (EFA/CFA).

On the issue of using individually single-item Likert-type items, Wanous et al. (1997) make the distinction between operationalizing psychological constructs (e.g. competence, event desirability) and measuring self-reported facts (e.g. years of education, work experience). They cite Sackett and

Larson (1990) as the earliest influential source attempting to justify conceptually the use of single-item measures in the cases of straightforward constructs. Such constructs are defined as sufficiently narrow and unambiguous (e.g. see Rossiter 2002; more on this in the next paragraph). For example, such a construct is often taken to be job satisfaction, whereas an example construct on the opposite end of this spectrum would be personality.

The strong tradition in academic marketing research of using multiple-item measures is usually associated with Churchill's (1979) influential article in the spirit of psychometric research on personality and aptitude of authors such as Guilford (1954) and Nunnally (1978). More recently, Rossiter (2002) picked up the baton from Sacket and Larson and theorized about the extenuating circumstances that exist under which the use of single-item measures should not be frowned upon. Such are the cases when the construct (e.g. desirability attitudes) can be categorized as doubly concrete, i.e. concrete attribute (e.g. likeability/desirability of) of a concrete singular object (e.g. uncertain event within a source). Concrete here is used in the sense that the attribute or object has consistent meaning across respondents, whereas an object is classified as singular only if it is perceived as lacking lower-level constituents or components. A very useful overview of the main arguments for and against the use of single-item measures can be found in Bergkivst and Rossiter (2007).

First, critics such as Churchill (1979) and Peter (1979) argue that multiple-item measures enable the computation of internal consistency indexes, i.e. establishing whether there is strong and positive correlation between the individual items. This is obviously not possible to compute for single-item measures. However, Bergkvist and Rossiter point to theory developed by the latter in 2002 which cautions against the use of such internal consistency checks without prior scale unidimensionality tests (e.g. see Cortina 1993) such as factor analysis or coefficient beta (e.g. see Revelle 1979). They further argue that the coefficient alpha (the dominant technique used by social scientists to test the reliability of scales; equivalent to taking the average of all possible split-halves Spearman-Brown-corrected reliability coefficients) is in fact only relevant for measuring attributes of constructs such as personality (provided the assumption of unidimensionality is satisfied, e.g. see Peters, 2014) and is much less meaningful for concrete (e.g. likeability) and formed (e.g. social class) attributes. Testing the reliability of single-item measures is especially problematic for two main reasons. First, there is no universally accepted way of computing reliability for single-item measures (alpha is not applicable) and, second, whatever method the researcher ends up using to successfully compute it, the resulting reliability score would inevitably be too low (Wanous and Hudy, 2001).

Later in this subsection I investigate the most popular approaches followed in the empirical literature. It is further argued that if predictive validity is equal between single- and multiple-item measures then reliability (internal consistency) is irrelevant for all practical purposes. According to Crutzen and Peters (2017), when it comes to assessing the validity of an experimental measure there are no hard rules. Establishing that the observed results indeed support the preferred interpretation necessitates a consistency check with past empirical studies and the relevant theory. Such a review of relevant empirical studies is the approach followed in this paper for both reliability and validity (to follow later in this subsection) due to the practical difficulties (i.e. the lack of adequate logistics and budget which necessitated the use of a shorter questionnaire administered to complete volunteers via digital means) of putting in place an experimental design which would allow for first-hand evidence to be obtained.

A second argument of the supporters of multiple-item measures is that single-item measures cannot capture all the different facets of the underlying construct of interest. Another nuance of this argument is that multiple-item measures provide a larger set of unique response patterns and total scores and are, therefore, more discriminating and contain more information. However, and once more pointing to the theoretical work of Rossiter (2002), the authors point out that multiple facets are not features of concrete and singular constructs and that the assumption of extra information contained in multiple-item measures is dependent on respondents' ability to make distinctions between the possible response categories. A further argument can be found in Rossiter (2002) who points out that many researchers lean on the work of Churchill (1979) in order to justify using multiple-item measures. However, Rossiter believes that Churchill made these claims in the context of ability-test theory, and hence such an argument is unfounded. In such a case it would be appropriate to use a multiple-item measures as ability difficult to capture with just one test question. Hence, he concludes that Churchill's paper can be of little usefulness to researchers who operationalize psychological constructs independent of respondent ability. A final oft-cited argument against the use of single-item measures for psychological constructs is the popularity of structural equation modelling which is a technique used to analyze underlying structural relationships.

From a theoretical point of view, Diamantopolous et al. (2012) summarize the key issues critics have raise a through conceptual illustration. Suppose that the measure x_1 is related to the unobserved variable η in the following manner:

$$x_1 = \lambda_1 \eta + \varepsilon_1$$

Lambda is the factor loading and epsilon is measurement error. The authors continue that there are two ways in which to interpret the measure. It can either be treated as unique or as representative. Both treatments offer theoretical and practical complications. For example, the former suggests that the measure can capture the construct in full, virtually making the two identical and indistinguishable from each other. This situation makes the construct devoid of significance and implications. For example, if the eliciting construct innovativeness is exhausted in full by the observable measure capturing the number of new products brought to market within a given time frame, then the former contributes little both semantically and practically making it redundant. In contrast, the latter assumes that the observable measure does not explain the complete meaning of the construct but rather that it is interchangeable yet sufficient as a standalone measure. This then gives birth to the non-trivial task of choosing the "best" candidate to represent the construct. Multiple such methods are discussed, all with their respective upsides and downsides, leaving the unfortunate practitioner with a lot of room for subjective judgement.

In contrast, Bergkvist and Rossiter believe that the arguments for the use of single-item measures mainly stem from practical reasons related to ease and cost of collecting primary data. They point out that the work of Rossiter (2002) is, at the time, the only serious attempt at theoretical justification for single-item measures. As alluded to in the previous paragraphs, in Rossiter's theoretical framework, in order to describe how it should be measured, a construct can be expressed in terms of its attribute ("dimension of judgment", and its components if any), object ("focal object being rated", including its constituents or components if any) and rater identity (pp. 10-11). This conceptual framework extends the work done by McGuire (1989) with a third aspect which is rater identity. It was mentioned earlier that the object part of a construct is defined in terms of two dimensions – singular/non-singular (collective or formed) and concrete/abstract. In this terminology, an object is classified as concrete if there is little disagreement between respondents regarding what the object constitutes, i.e. they all describe it in an identical way. Similarly, for an object to be classified as singular, respondents must agree that what they are asked to rate is indeed only one object, or if multiple then at least homogenous to a reasonable degree, and that it is not made up of different components or constituents. A concrete object is always singular, and vice-versa. Additionally, abstract-collective objects have different meaning across respondents but *collectively* form a higher-level object, and hence the sub-objects are called constituents and measured unidimensionally according to a single attribute.

In contrast, abstract-formed objects also have different meaning across respondents but in a sense that respondents see the object as having different *components* which make up the meaning of the construct, so that it should not necessarily be measured unidimensionally. The authors point out that the former answer the question what the object includes whereas the latter answer the question what it means. The attribute part of the construct can be either concrete (singular) or abstract (formed or eliciting). In the case of the latter, respondents attach different meaning to the attribute which is either made up of different components which together add up to realize it (formed) or is causally linked to the components which represent the attribute's outer manifestation (elicited) or "*proximal consequences*" (p. 25). In essence, what distinguishes these two is the direction of the causal effect. In the case of formed, it is the components which *cause* the attribute, whereas in the case of eliciting it is the attribute which *causes* the components (similar to the latent variable approach). Two examples given in his work are service quality and service emphasis, respectively. Once again, concrete refers to the idea that respondents understand and agree on what dimension they are asked to judge the object and that it is indeed a single characteristic.

When it comes to the empirical arguments against the use of multiple-item measures (and in favour of single-item measures), the strongest case applies for abstract attribute measurement involving a selection of synonymous items (questions). In such cases there is an elevated danger that increasing the number of items in the battery also increases the probability of including non-synonymous items and, with it, the probability of respondents answering a non-synonymous and synonymous item in identical way (hurting content validity). Additionally, within the domain of doubly concrete constructs, Drolet and Morrison (2001) show that high error correlation renders pointless the inclusion of many additional items. For example, according to their calculations, the extra information provided by an infinite number of items with 0.60 error correlation compared to two items with 0.60 correlation is equivalent to 0.42 additional independent items.

Another study which tests the reasonableness of arguments against the use of single-item measures is the one by Wanous et al. (1997). They conduct a meta-analysis of research on the correlations between individual single-item Likert-type items capturing job satisfaction (a theoretical construct for which the use of single-item measures is well established, the authors note, and is often treated as doubly-concrete in Rossiter's terminology) attitudes and Likert scales (defined earlier as the aggregation of multiple such single-item Likert-type responses). They find on average a correlation greater than 0.60 for both observed and corrected means between single-item measures and Likert scales. They also note that correlation is highest for scales focused on overall job satisfaction rather than made up of single-item responses focused on different aspects of overall job satisfaction. They interpret their findings as evidence favoring the use of single-item measures provided certain conditions exist (such as the construct being narrow and unambiguous). For example, they also believe that single-item measures should be used in place of scales in cases when aggregating the individual responses eliminates meaningful variability.

This issue also applies to the present paper as illustrated in subsection 3.2 with the difficulty of interpreting global/aggregated desirability ratings. Scarpello and Campbell (1983) support such findings and also distinguish on conceptual level single-item measures of global preferences or attitudes from multiple-item measures of different aspects of the studied construct. They believe that the former are useful in their own right and may contain information not typically found in multiple-item measures which themselves may suffer from the potential omission of relevant aspects of the construct that influence preferences or attitudes. Hence, they conclude that it may often be the case that the information captured by the global question is more than the sum of the distinct aspects which are chosen to be tested. Empirical support for singe-item measures can also be found in other scientific domains. For example, McKenzie and Marks (1999) test the validity and reliability of single-item assessment of mood compared to more time-consuming measures. Their results indicate a 0.78

correlation between the single-item measure of depressed mood and the factor scores of the latent construct underlying the multi-item scale. The single-item measure also generates a very high inter-rater agreement as the scores of patients and clinicians have 0.91 correlation. They conclude that such a measure can save a great amount of time for both patients and clinicians, and that it is a valid and reliable rough guide for mood.

It may also be useful to spare a thought about the method used by many of the studies cited in the previous paragraph to establish their main result (i.e. that under certain conditions, single-item measures can be as reliable, and sometimes even more, as Likert scales). Earlier in this subsection it was mentioned that there is no universally agreed upon method of establishing the reliability of single-item measures. While this is true, the one method preferred by many researchers is the correction for attenuation formula introduced in Nunnally and Bernstein (1994).

$$\hat{r}_{xy} = r_{xy} / \sqrt{r_{xx} r_{yy}}$$

This method gathered popularity after it was adapted in Wanous et al. (1997) to prove that it can be used to solve for minimum reliability of single-item measures. According to Nunnally and Bernstein, the method requires two measures (x and y) of distinct theoretical constructs (e.g. desirability and over-optimism) but it also works for two measures of the same construct (usually one single- and one multiple-item measure). The leftmost term stands for the assumed underlying correlation. A good illustration of this approach and an extension of the original study of Wanous et al. is the work by Dolbier et al. (2004) who compare (attenuation-corrected) reliability between alternative measures (multi- and single-item) of job satisfaction. They create two measures of job satisfaction out of one 16-item measures – one with a single-item overall satisfaction measure and one aggregating the rest of the 15 remaining items. Their results show 0.73 minimum single-item reliability (when the underlying correlation between the two measures is assumed to be perfect), slightly higher than reported in Wanous et al. They also replicate other measures (in addition to the multiple-item job satisfaction measure) to test concurrent, divergent and convergent validity scores. The single-item measure was positively and significantly correlated with the job satisfaction scale and other scales measuring perceived supervisor and coworker support (providing evidence for convergent validity), and negatively and significantly correlated with measures of work stress – providing evidence for divergent validity.

Extending this line of research, Bergkvist and Rossiter (2007) conduct a study comparing the validity of single-item and multiple-item measures of marketing constructs (i.e. psychological constructs rather than factual). They employ a within-subject design to test brand and ad attitudes. The results from their analysis lead them to conclude that single-item measures of both constructs are equally valid to multiple-item measures as per two methods of testing predictive validity, namely bivariate correlation and multivariate regression r-squared statistic. However, they are careful to caution that their study has implications that are limited to the domain of doubly concrete constructs (Rossiter 2002; more on this in later paragraphs), and hence generalizations are not advisable.

Spurred by Bergkvist and Rossiter's (2007) study, Diamantopoulos et al. (2012) contribute to the empirical literature investigating the conditions under which predictive validity of single-item measures does not differ much from that of multiple-item measures. They replicate the former's research and conduct simulation analysis on the relative importance of different design characteristics for the predictive validity of a measure. In the process they establish the following guidelines (see Figure 1) for researchers to use when deciding between single- and multiple-item measures. First, for small sample sizes (e.g. less than 50) both types of measures show identical predictive validity, and hence it is reasonable to prefer the former in this case. Second and third rules state that when cross-item correlations are low (e.g. around 0.30) or when the items are highly homogenous semantically (e.g. high inter-item or Cronbach's alpha, coefficients suggesting strong internal consistency), then it is also preferable to employ single-item measures. Using many semantically close items (e.g. synonyms) poses the danger of damaging content validity (e.g. Rossiter, 2002; Drolet and Morrison, 2001). In conclusion, Diamantopoulos et al. close that single-item measures seem to be most appropriate in the cases of exploratory research, as it often matches well the rules set out above, prioritizing mapping out the main present effects rather than providing a detailed treatment of underlying relationships and effect magnitudes. In this context it makes sense to ask one "global" question as it encourages respondents to consider the issue from all perspectives which are relevant to them instead of attempting to discriminate between different factors and their relative importance. The authors also echo Rossiter's (2002) recommendations that in such circumstances it is crucial that the single item is specific and clear to each respondent in a consistent manner across respondents (i.e. construct is doubly concrete, in Rossiter's language).

Further support for the use of single-item measures can be found in prior empirical literature which studies similar theoretical concepts. For example, Massey, Simmons, Armor 2011 operationalize their "preferences" (in their study - favorite NFL team and degree of likeability of different NFL teams) and "familiarity" (how much subjects know of each NFL team) constructs with single-item, 5-point (answer options) measures. Morewedge et al. (2018) sample their respondents pre-existing inclinations towards certain resolutions via a closed-end measure with several alternative (mutually exclusive and collectively exhaustive) answer options. Olsen (1997) measures event desirability with one single-item, 7-point Likert-type question per event, asking his subjects to choose between alternative answer options on a unipolar spectrum ranging from "1 = very desirable" to "7 = very undesirable". Price (2000) extends the line of research under the umbrella of the minimal group paradigm – a set of studies related by the use (social) group affiliations as the main factor influencing their respondents' judgements or choices. It is argued that such experimental design replicates very closely the setting in key field studies boasting effective event desirability manipulations. Heath and Tversky (1991) operationalize competence with a single question asking respondents to rate their confidence on a spectrum ranging from pure guessing to certainty. Similarly, in another experiment in their paper, they also ask subjects to rate their knowledge of football games using a single-item, five-point Likert scoring or even a binary high-low knowledge measure.

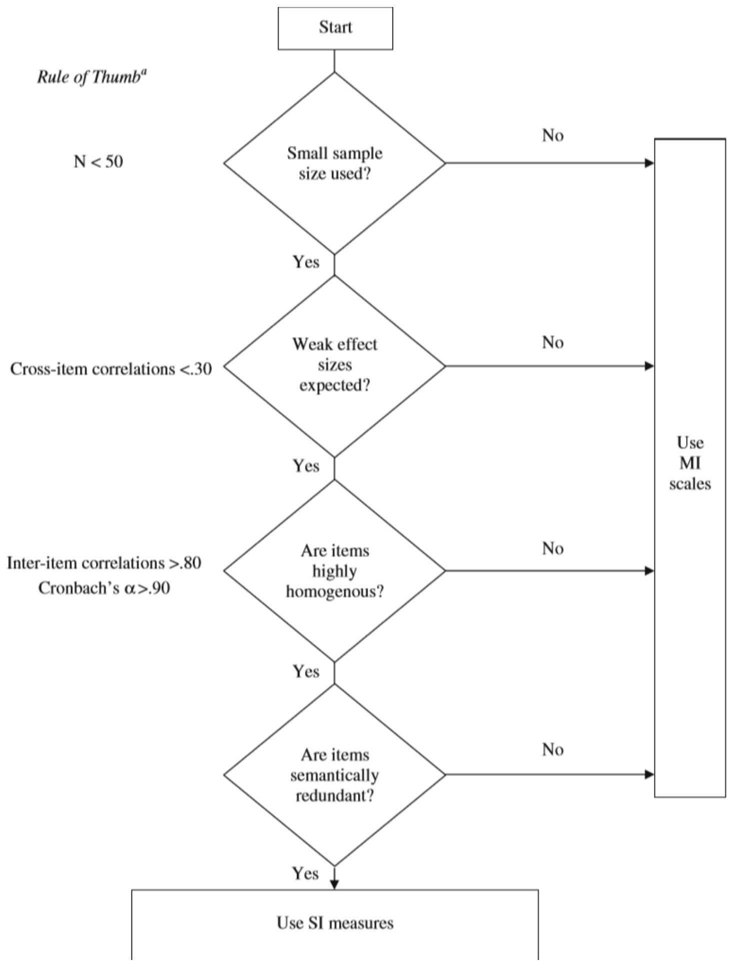


Figure 1. Adapted from Diamantopoulos et al. (2002)

Application to the present study:

The analysis in this subsection so far paints an ambiguous picture as to the correct way of treating the main independent variables in the present study – event desirability and competence. For example, Rossiter himself admits that for an attribute such as attitude (here desirability), it all depends on how the construct (here attitude or equivalent to attribute depending on the theoretical treatment) is conceptualized, and hence there are no strict rules whether it should be classified as concrete, formed or eliciting. For example, often, and this also applies to the present paper, attitude is conceived of as an overall evaluation. Rossiter (2002) observes that if the purpose of the study is attitude formation, then this overall evaluation can reasonably be classified as formed attribute. As mentioned, in Rossiter's language, a formed attribute of a construct is such that its components "*make it appear*" (p. 20). Rossiter also identifies several properties of the components of the formed attribute. First, the components must themselves be attributes, and if they are first-order attributes (that is they are not also formed attributes) then they must be concrete as well. Each first-order component can then be measured with a single item. Second, expert judgement determines the most representative components to be included. Third, all these main components must be measured with separate items. Fourth, a formed attribute (and hence the generated scale) is never unidimensional, and therefore proving a single common factor using factor analysis techniques or high internal consistency using Cronbach's Alpha is irrelevant.

In contrast, studies on established attitudes have more reasons to treat the overall evaluation of an object as a concrete attribute of a construct. Yet a third line of research exists which posits that the overall evaluation is never a concrete characteristic because attitude is a multi-component (second-order eliciting) attribute. In this view, the overall evaluation of an object always consists of a cognitive and an affective part, which are both *elicited* by and an outer manifestation of the inner (latent) attribute - attitude. In this treatment, it is the latent attribute of the construct which causes the observed responses to the items. Hence, the (second-order) eliciting attribute should be unidimensional and this applies to the components (first-order eliciting attributes) too. Moreover, it is conceptually a necessary condition for the component scores to be unidimensional since in theory it is the latent attribute which is causing the item responses. Rossiter recommends three to five items, assuming three to five components and a first-order eliciting attribute. He also recommends that the separate items measuring the eliciting attributes are concrete and distinct activities. Hence, reasonable arguments could be made in favor of the argument for treating the attribute desirability in the present study as eliciting, if we indeed assume the presence of such a two-system evaluation.

As discussed in section 2.2.2, authors such as Lench (2009) support the affective interpretation of desirability and such a two-system treatment is also in unison with the ideas presented in section 3.2 that desirability evaluations may have a rational (i.e. cognitive) and irrational (i.e. affective or what this study refers to as systematic influence stemming from the source level) part. Hence, the presence of a two-system evaluation would offer support that the attribute of the construct desirability attitude may indeed be eliciting, and perhaps also that a systematic influence of "source desirability" over the event-specific evaluations may be present. However, to remind the reader, this paper does not attempt to test whether the two-system theory (or any other psychological source theory for that matter) may be plausible or not.

Although investigating such questions may potentially lead to powerful insights into the essence of desirability bias, it is nevertheless beyond the scope of this study. Arguably, the experimental design does not allow for establishing such a result as the attribute is treated as concrete, and hence only three questions overall are used (one for each single event). In fact, the for the purposes of the present paper the conceptual distinction between psychological construct (e.g. attitude) and the actual measure (e.g. event desirability) – the attribute of the object or the characteristic on which the object of the construct is evaluated - is meaningless. In contrast, testing such conjecture would necessitate two sets of at least three (sets of three) questions per first-order eliciting attribute (per constituent of the abstract collective object) – cognitive and affective component – and factor analysis to prove the unidimensionality of each (first-order) component. Factor analyzing a concrete attribute is also expected to yield one common factor explaining most of the variability. However, since only three questions are used overall, any evidence supporting the hypothesized two-system nature of desirability attitudes may be left undiscovered.

I would argue, based on the arguments presented in this subsection and based on my analysis of the main independent variables in subsection 3.2, that the competence and desirability constructs differ not only in the attribute which is measured but also in the focal object. For the former, I would further argue that the object is the source itself as rating the separate events which make up the source is a redundant and pointless exercise in this case. In contrast, for the latter, in order to allow potentially for the rational and irrational part of the desirability evaluation, we need to break down the source of uncertainty into its (mutually exclusive and collectively exhaustive) constituents, and thus treat the object here as abstract collective (effectively splitting the sample in three) but its constituents – as concrete. This classification appears to be the most appropriate since the object of the construct does not rely on its constituents (all possible single and composite events) to derive its meaning – a key characteristic of abstract formed objects. The definition of the abstract collective object is unambiguous, and it poses certain restriction on what represents a valid constituent.

A single-item measurement approach is thus followed in the present study for competence at the construct (here equivalent to object level) level while for desirability - at the constituent level. Hence, the individual responses will be used in their raw format as it is deemed unnecessary and inappropriate to use involved aggregation procedures (SEM) for the creation of a scale. This approach is also taken out of practical considerations. However, given the literature review conducted in this subsection, both theoretical and empirical justifications can also be provided. Moreover, as discussed before, a similar treatment of related constructs is found in the extant literature. Due to fact that the relevant experimental literature is still very thin as of the date of writing this paper, there is a lack of competing treatments of event (un)desirability to support the alternative hypothesis that it should not be operationalized with a single-item measure (for example because it is unreasonable to treat the attribute as concrete). However, it is also important to recognize the weakness of Rossiter's (2002) conceptual framework on which the treatment of desirability in the present study is modelled – and that is the practical difficulty of testing the conjecture put forward in his paper even via means of expert judgment (the only method approved by Rossiter). Although this proposition cannot be tested directly, the recommendations from other studies discussed here appear to be supportive of using single-item measures.

4.1.2. CHOOSING BETWEEN PARAMETRIC AND NON-PARAMETRIC TESTS

Background information:

The question about the most appropriate way to classify data generated by Likert-type response formats has long been a ground for burning debate between scholars (e.g. Allen and Seaman, 2007; Carifio and Perla, 2004; Jamieson, 2004). As initially defined in Stevens (1946), data generally fall under four levels of measurement depending on the amount of information they provide – nominal (or categorical), ordinal (or ordered categorical), interval (equi-distant categories), and ratio (zero has meaning; in ascending order of information richness). In a nutshell, the issue concerns the reasonableness of treating Likert-type data as interval as opposed to ordinal. According to Allen and Seaman, the underlying reason for the existence of this polemic is to ascertain whether the use of parametric tests is allowed with Likert-type data or researchers have to resort to nonparametric techniques which often have less statistical power (greater probability of making type II error) and are less straightforward to interpret. As the authors explain further, if the most appropriate classification of such data is ordinal, then hypotheses cannot be defined and tested in terms of, for example, means and standard deviations as such parameters rely on the assumption of normal distribution. Instead, distribution-free statistical methods – relying on median, rank and frequencies - should be used to describe the data and test the hypotheses. The significance of this debate is magnified by the great popularity of Likert-type questionnaires and surveys in applied social science research (de Winter and Dodou, 2010). There is an abundance of consistent evidence that parametric tests such as the Student's t-test are more powerful for data with normal distribution and equal variances (for references see de Winter and Dodou, 2010 pp. 1-2). On the other hand, nonparametric tests, such as Mann-Whitney-Wilcoxon, hold a power advantage for non-normal or data with outliers (again, see de Winter and Dodou, 2010 p. 2).

One of the main arguments that opposers of the idea of treating Likert-type data as interval cite is that verbal categories rarely have labels which can successfully be proved to generate identical steps in feelings intensity (Jamieson, 2004). Often, critics believe, these verbal labels will create consecutive categories with a psychological distance between them that is perceived to be unequal by respondents. Hence, this would violate the key assumption of the interval level of measurement as defined by Stevens and can potentially make the output from inferential statistics highly misleading. It is important to note that this argument is significantly less relevant for Likert scales with numerically-labeled categories. Second, Carifio and Perla (2008) posit that the origins of the arguments against treating such data as interval can be traced back to Stevens (1946). They argue that Stevens wrongly rejected the conjecture about the so called "emergent properties" of scales (p. 1150). This refers to the idea of the scale having properties which are different to those of the items which comprise it. Stevens apparently believed that if Likert-type items are individually ordinal, then collectively (Likert scales) they cannot be anything more than that. However, Carifio and Perla argue that Stevens misused (and possibly misunderstood too) what Rensis Likert originally meant by the term “scales”.

However, Carifio and Perla (2008) turn to the weight of empirical evidence and claim with strong conviction that it clearly tips the balance in favor of classifying Likert-scales data as interval. For example, Norman (2010), although in agreement with Jamieson (2004) that equi-distance between consecutive categories is often a questionable assumption and that choosing the inappropriate type of test increases the probability of making incorrect inferences, maintains that what most researchers usually fail to appreciate is the robustness of different statistics to certain degrees of violation of their key assumptions. Norman goes on to investigate the impact of three of the most hotly debated violations of the key assumptions of parametric methods – small sample size, non-normally distributed data, and ordinal level of measurement – and concludes that “parametric methods are incredibly versatile, powerful and comprehensive” (p. 3).

Therefore “intervalists” (to borrow the term from Carifio and Perla) would wholeheartedly agree with Rossiter (2002) who confidently states that researchers gathering primary data via means of Likert-type items who have ensured that their questionnaire possesses a valid psychological zero, provides for a separate “no opinion/don’t know” answer option signifying lack of awareness (as opposed to awareness but neutrality, i.e. the oft-used “neither agree nor disagree” category), has clear minimum and maximum intensity categories, and employs answer option labels making the categories appear (psychologically) equi-distant (more relevant for verbally-labeled categories), have solid grounds for arguing that the resulting data approximates at least interval if not ratio structure. Rhemtulla et al. (2012) adds to that list an additional requirement: the presence of no fewer than five categories. Their simulation shows that for variables with five to seven answer options robust continuous methods (maximum likelihood) for confirmatory factor analysis performed as well as robust categorical methods (categorical least squares).

Another important consideration for the choice between parametric and nonparametric tests is, of course, the sample size, though there are again no hard and fast rules (de Winter and Dodou, 2010). Jamieson (2004) also concedes the possibility that indeed the choice between parametric and nonparametric tests for Likert-type data may be better made based on sample size and distribution rather than on the level of measurement. The author advises caution even when a researcher is certain her data come in the interval format and recommends considering the sample size and confirming whether the distribution is normal. This view appears to be supported by Knapp (1990) who observes that Stevens (or anyone after him) did not specify exact rules for determining the right level of a measurement for a given dataset.

Unfortunately, scholars cannot seem to agree even on this issue. Norman (2010) contends that the assumption of a sample size floor for parametric procedures such as ANOVA is a pure fabrication and that such a rule simply does not exist in the statistics textbooks. Although, he does concede that small sample size is relevant from non-statistical perspectives, e.g. external validity. Moreover, Norman also strongly disagrees with the other frequently cited obstacle to the use of parametric tests for below-interval level data – that of normal distribution. He reminds his readers that the assumption of normality does not concern the distribution of the observations themselves but that of their means, and that the Central Limit Theorem, as well as many authoritative empirical studies on robustness, proves that this holds even for very small sample sizes irrespective of the distribution of the underlying raw data.

A downside of working with parametric procedures in the case of multiple-item measures is the potential for losing meaningful information in the data aggregation process. As mentioned in subsections 3.2 and 4.1.1, when the underlying data come in a limited set of distinct categories ("ordinalness" is an attribute of the data rather than of the labels, as perhaps Allen and Seaman would put it), inspecting a single parameter such as the mean will in some cases obscure important variability. This is especially true in cases where data are clustered around both ends of the response spectrum with few observations in the middle. Inspecting the frequencies will instead paint a different picture.

In his seminal paper, Likert (1932) makes the implicit assumption that there always exists a latent (or "natural") variable underlying the respondents' attitudes, and hence, as Clason and Dormody (1994) put it, "*[e]ach Likert-type item provides a discrete approximation of the continuous latent variable*" (p. 32). This concept was already touched upon at the very beginning of subsection 4.1.1. Therefore, the authors stress that any researcher who fails to acknowledge the discrete nature of each individual response to a Likert-type item is committing an error on a conceptual and possibly inferential level as well (refer back to the example discussed in subsection 3.2 and 4.1.1 on the potential loss of meaningful variability if a discrete item is analyzed as continuous). Furthermore, the authors find it difficult to imagine how individual Likert-type items may generate data approximately following normal distribution due to the common problems of floor and ceiling effects, producing skewed distribution instead. When the artificial (and arguably arbitrary) limits imposed make respondents choose either of the extreme answer options, the, the sample size, rather than respondent attitudes, may become the driver behind the observed responses. The authors also advise against deciding on the appropriate inferential methods ex-post, e.g. after ascertaining the distribution of the data via a Kolmogorov-Smirnov or other procedure (e.g. Anderson-Darling, Shapiro-Wilk). They agree with others who express concerns that the subjectivity in drawing the line between data that are sufficiently close to normal distribution (if that has any significance at all) and such that are not (e.g. due to the effects of sample size on existing test of distribution) may lead researchers to make decisions about the types of tests to use based on what fits their personal agenda.

Application to the present study:

It should be noted that all the discussion in this subsection up until the last paragraph concerns Likert scales, or the summated individual Likert-type items. And while the evidence in favor of treating Likert scales data as interval may appear somewhat convincing, it would not be acceptable to simply assume that the same should automatically hold for individual Likert-type items. Given the literature review in this subsection, it is concluded that it is safer to employ non-parametric procedures mainly due to the use of single Likert-type items. Moreover, it can also reasonably be concluded that such may yield greater statistical power in this case due to the non-normal distribution of the data and the relatively small sample size.

4.2. EXAMINING THE EFFECTIVENESS OF THE CONDITIONS

First, I attempt to prove whether a certain response pattern exists within the different conditions, in order to then proceed to ascertaining whether the three experimental groups differ significantly from each other. I examine the frequency distribution of the desirability scores for each condition and use nonparametric methods to achieve these goals. I also attempt to determine whether the different conditions produce responses in the hypothesized directions (i.e. of predicted sign). Recall, from subsection 3.2 that the two treatment conditions were chosen with the assumption that each describes a source of uncertainty of contrasting desirability to the other. That is to mean the neutral condition should generate observations which cluster around zero, whereas the desirability and undesirability treatments should produce responses predominantly in the positive and negative end of the answer choice spectrum, respectively.

Hence, a hierarchy of responses can be established, if the data are pooled together, in which observations from the desirability condition should consistently receive the highest ranking, followed by the neutral and then the undesirability condition. It would also be interesting to see what the frequency count is of perceived competence self-evaluation scores between conditions. Do subjects perceive themselves equally knowledgeable or ignorant about all three sources of uncertainty or are there significant differences? Graphical representations of the data enable us to eyeball some of these propositions. All the methods discussed in this and the next subsection will adhere to the conclusion from the previous subsection (4.1.2) that the independent variables constructed via individual Likert-type items from the questionnaire are of ordinal nature, and hence parameters such as means and standard deviations are not appropriate.

A Chi-squared test can be performed to check whether there is indeed a trend in the (un)desirability ratings provided by respondents or whether their answers follow a purely arbitrary path. If that were the case, then out of the *y* total observations (respondent-answers, i.e. x respondents times 3 answers per respondent), each of the 7 categories (or ratings) should receive an expected frequency of *y* divided by 7. If the null is rejected (i.e. the observed frequency is significantly different from the expected frequency), then a pattern obviously exists. Unfortunately, explicitly testing whether the observation from the control condition are zero on average, and non-zero in the treatment conditions, would involve parametric procedures. One potential solution involving a nonparametric alternative could be the Binomial test. This test would check whether event desirability ratings drawn at random from any of the three conditions follow the predicted frequency. For example, defining the negative end of the response spectrum as X=1 and the positive end as X=0, the test can determine whether a rating drawn randomly from the undesirability condition has 50% probability to be from either end or the alternative being greater than 50% of being from the hypothesized end of the response spectrum (i.e. left). Define the number of trials to be the total number of ratings, the number of successes to be the ratings which are within the [-3,-1] closed interval, the null hypothesis to be that the disjunctive event probability is equal to 3/7, i.e. the resulting test statistic would give the probability of observing the given number of successes with the null still not being false, and the alternative that it is greater than 3/7, i.e. the observed count of successes could not have materialized if the probability of occurring was indeed equal to 3/7. In Stata this is invoked via the code "*bitest varname== #p [if] [in] [weight], [detail]*" for the immediate form of the binomial probability test and "*bitesti #N #succ #p, [detail]*", where *varname* includes only 1's (i.e. "successes") and 0's.

Subsequently, I attempt to answer the question if the control (desirability-neutral) condition is significantly different from the treatments using the Kruskal-Wallis test. Formally, this test provides evidence if the observations in all the three conditions came from the same underlying population. If the null hypothesis is not rejected (i.e. no evidence that the three groups of observations do not come from the same population), then I can proceed to pool them together and form one big treatment condition which is hypothesized to be different from the desirability-neutral condition. This can be tested using the two-sample Mann-Whitney U test. Finally, I conduct a Jonckheere test. This test is used to determine whether the scores follow the predicted rank order, i.e. observations from the undesirability condition consistently receive the lowest rank, followed by the scores in the control condition, followed by the scores in the desirability condition.

As a form of robustness check for the results, I employ the multinomial logistic regression analysis with the ratings (e.g. desirability scores) as explained variables and the conditions as explanatory variables (including control variables). To make this method work, I group the separate scores into a binary variable. For example, to test the hypothesis that being allocated to the undesirability condition is associated with greater probability of providing a negative score, I code all negative categories as 1 and all non-negative answer options as zero. And vice-versa, to test the hypothesis that being allocated to the desirability condition is associated with greater probability of providing a positive score, I group all positive categories and code them as 1 while all non-positive – as 0. This method also insures against the case in which some categories possess very low frequency distribution.

4.3. EXAMINING THE PRIMARY HYPOTHESES

The appropriate non-parametric procedures for testing the main hypotheses are similar to those discussed in the previous subsection. Again, this is the preferred method of analysis as the main independent variables (competence and desirability) come from individually-analyzed Likert-type items while the main dependent variables (a-aversion/insensitivity), although not strictly discrete, come from a range with upper and lower bounds, and hence cannot assume any value. Given the level of measurement of the data and the hypothesized presence of concomitant effects of perceived competence on ambiguity attitudes, a slight complication to the analysis can be anticipated. The covariate is of ordinal character and nonparametric methods that can account for ordinal covariates are not widely used.

Therefore, two general routes can be taken to test the main hypotheses. First, it may be useful to determine in advance whether in this database the covariate (i.e. competence) has any impact at all on the main explained variables (i.e. the ambiguity indexes). Therefore, to circumvent the need to resort to more obscure statistical methods, I flip the model around, so that ordinal techniques like multinomial and binary dependent variable regressions can be used. Establishing the absence of any significant association would be one factor used to determine if the variable should be retained for further analyses. For example, if it has very low and insignificant correlation with ambiguity-attitudes and the experimental group variables, to determine whether the three experimental groups differ in terms of their impact on the main dependent variables, the case could be made again for the use of the Kruskal-Wallis test without any regard for the conjectured covariate.

The second approach, although not widely used, involves considering the use of a nonparametric alternative to the widely employed parametric ANCOVA or multiple regression analyses. As pointed out in Quade (1967), the classical parametric analysis of covariance (ANCOVA) assumes under the null a normal distribution of the dependent variable given the covariate, a linear relationship between the two, and a variance that is independent of the latter. Instead, the rank method he proposes (RANCOVA) avoids all these requirements and replaces them with the assumption that the distribution of the covariate is independent of the group. The test retains a similar interpretation as that of the original, i.e. the probability that an arbitrarily-drawn observation from one group will exceed that of another. Quade (1967) opines that a related route often taken by researchers, when their data do not meet the assumptions required to perform parametric testing, is to search for suitable methods of transforming that data (arguing that his approach should be distinguished from this). However, he warns that those are difficult to find and, in any case, subsequently present further complications in terms of computation and interpretation. An example of such method, in the spirit of the Quade nonparametric test but also enabling us to test the hypothesis for equal slopes between groups, involves rank-transforming the raw data and conducting regular ANCOVA on the resulting ranks (Conover and Iman, 1982). The authors show that their method yields robust results, identical to those from the Quade test.

Finally, the Wilcoxon (signed rank) test could be used to answer within-subject design questions such as whether there is any significant difference between ambiguity attitudes in the Ellsberg experiment and ambiguity attitudes in the control/treatment conditions. Formally, this is a two-sample method for determining whether two paired samples come from the same population.

5. RESULTS FROM STATISTICAL TESTS

Section 5 proceeds as follows – 5.1 and 5.3 describe the data using various nonparametric descriptive statistics and present mostly informal tests connected to the different hypotheses in the form of visual evidence, while 5.2 and 5.4 present results from formal quantitative tests of the secondary and primary tentative hypotheses of interest, respectively. Subsections 5.6 and 5.7 summarize and offer an interpretation of the key results, in that order.

5.1. DESCRIPTIVE STATISTICS – EXPLANATORY VARIABLES

Figure 2 presents the pooled perceived competence scores from across the three conditions. Relatively more respondents indicated minimal or complete lack of knowledge on a given topic with the smallest bucket being that of complete perceived competence (tied with the spectrum midpoint). Given the psychological appeal of the midpoint (e.g. Krosnick et al., 1996), it is interesting to observe that respondents were reluctant to choose the middle answer option on this item.

It may also be insightful to examine how these competence ratings are spread out across the three experimental groups – Figure 3 presents this. The number codes on the table identify the condition according to the following order (without any significance attached in this case): 0 stands for the control group, while 1 stands for the undesirability treatment.

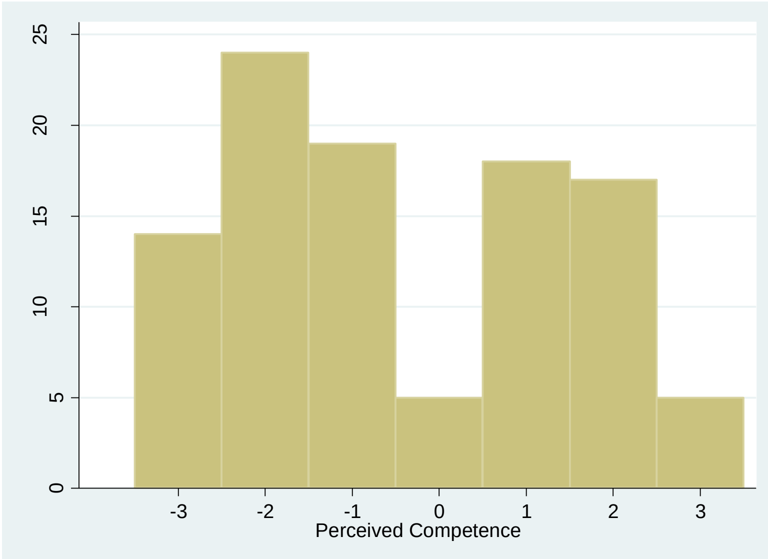


Figure 2: Histogram, Competence

The latter attracted the smallest number of negative competence scores suggesting greatest familiarity with the matter of job automation compared to the other two topics. Relatively speaking, the former possesses the largest imbalance, which also happens to be in the direction of the negative end of the choice spectrum. Formally testing these differences is important in order to ascertain that the RANCOVA assumption of orthogonality between the covariate and the experimental groups holds.

To do this, I collapse the 7-category perceived competence answer spectrum into a 3-category one because 2 answer options generated very low frequencies (as can be seen from Figure 2). I then compute the Spearman correlation coefficient (for it is free from distributional assumptions) and test the null hypothesis that the perceived competence ratings are independent from the experimental groups. The resulting test statistic (see Figure 4) is very large which provides very weak evidence against the null. Additionally, we see from Figure 5 that the chi-squared test for independence similarly fails to reject the null hypothesis that the two variables are independent.

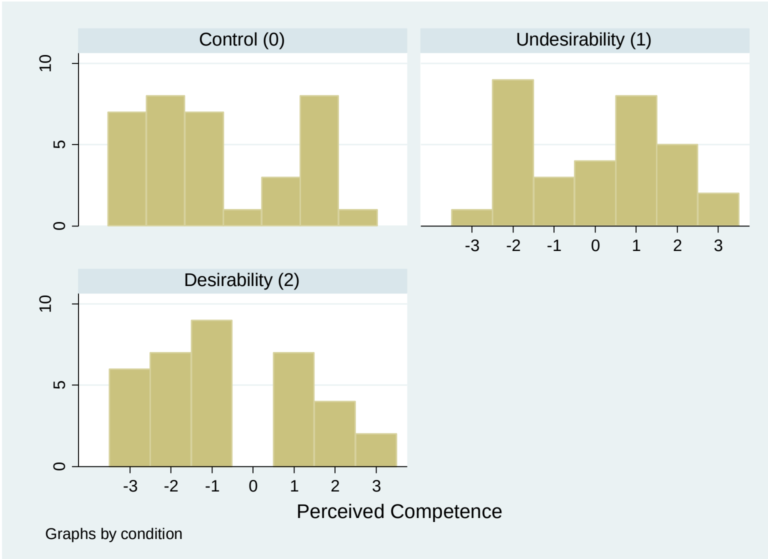


Figure 3: Histogram, Competence by Condition

```
. tab comp1_1

    comp1_1 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -3 |         14       13.73       13.73
         -2 |         24       23.53       37.25
         -1 |         19       18.63       55.88
          0 |          5        4.90       60.78
          1 |         18       17.65       78.43
          2 |         17       16.67       95.10
          3 |          5        4.90      100.00
------------+-----------------------------------
      Total |        102      100.00

. recode comp1_1 (-3/-2 = 0) (-1/1 = 1) (2/3 = 2), generate(competence)
(67 differences between comp1_1 and competence)

. tab competence

  RECODE of |
    comp1_1 |
  (comp1_1) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         38       37.25       37.25
          1 |         42       41.18       78.43
          2 |         22       21.57      100.00
------------+-----------------------------------
      Total |        102      100.00
```

```
. spearman condition competence

 Number of obs =      102
Spearman's rho =       -0.0069

Test of Ho: condition and competence are independent
    Prob > |t| =        0.9451
```

Figure 4: Spearman's Rho, Condition-Competence

```
. tab condition competence, chi2

           |    RECODE of comp1_1 (comp1_1)
 condition |         0          1          2 |     Total
-----------+---------------------------------+----------
         0 |        15         11          9 |        35 
         1 |        10         15          7 |        32 
         2 |        13         16          6 |        35 
-----------+---------------------------------+----------
     Total |        38         42         22 |       102 

          Pearson chi2(4) =   2.4379   Pr = 0.656
```

Figure 5: Chi-Squared, Condition-Competence

Similarly, I present the (un)desirability scores from the 3 groups per condition. Interestingly, the positive desirability scores have been chosen more frequently overall. Out of the 102 respondents (and 306 respondent-evaluations), 55 have chosen only non-negative answer options and of those 28 selected only positive. Additionally, there were 28 respondents who chose only non-positive answers with 5 of those providing only negative evaluations. This indicates that the undesirability treatment was not as successful in encouraging responses in certain direction since it was hypothesized a priori in that it would mainly generate negative scores. Hence, Figure 6, Figure 7 and Figure 8 represent the first (informal) test whether the treatments were successful in encouraging responses in certain directions. Eyeballing the frequency distributions, we immediately notice that the stimulus in the control condition produced almost exclusively neutral desirability evaluations as hypothesized. As for the two treatments, although there appears to be a pattern of the hypothesized shape, this is most pronounced on the final question and, overall, it does not show consistently across all questions. As a result, the two groups have generated what appear overall to be data of similar shape. Nevertheless, I proceed to conduct formal test for differences in responses dependent on the group the subject was allocated to.

As explained in subsection 4.1, for this purpose I employ the binomial test. First, I examine the evidence for and against the proposition that the number of respondents who have indicated a desirability evaluation of 0 in the control condition is in fact not due to chance or indeed cannot be generated by a random mechanism. In this case, the null hypothesis states that the probability of observing such a number of successes (responses of 0 on the choice spectrum), denoted by k in the Stata output below (Figure 9), is in fact not significantly different from 1/7 – the expectation based on the lack of any response patterns. I use Stata's "*recode*" command to replace all responses of 0 with 1 (i.e. success) and everything else - with 0. The actual observed answers which may be evidence against the null are 17 whereas the expectation is 5, in the case of the first (un)desirability Likert-type item in the control group. As it turns out, the probability of observing a value higher than k with the null still not being false is essentially zero, implying that we need to reject the null and conclude that the number of respondents answering with zero on this question is not a coincidence. The conclusion is identical for the other two items or questions in this condition.

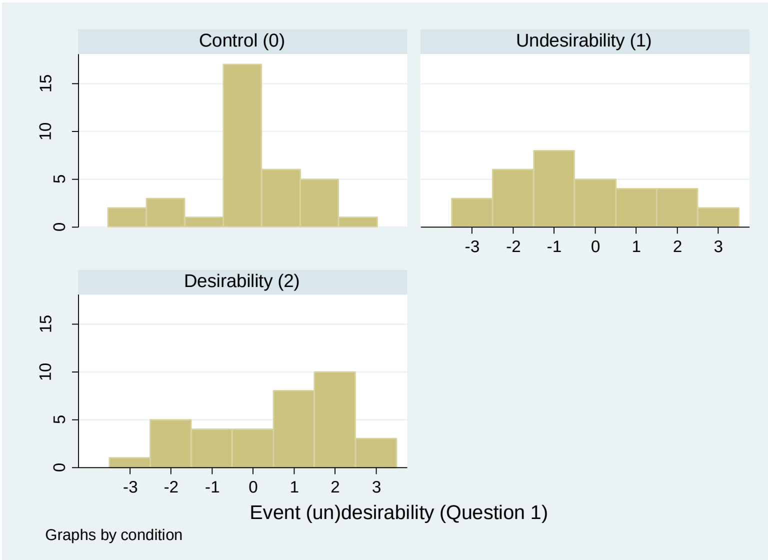


Figure 6: (Un)desirability scores by Condition, Event 1

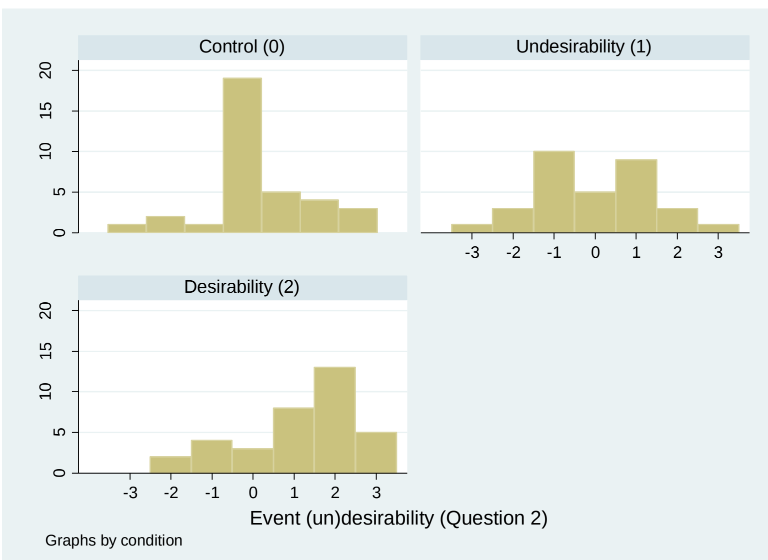


Figure 7: (Un)desirability scores by Condition, Event 2

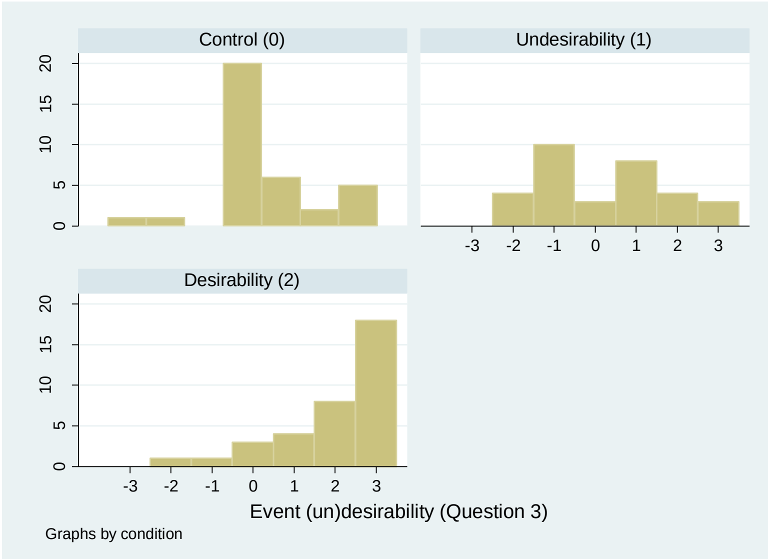


Figure 8: (Un)desirability scores by Condition, Event 3

Figure 10 presents the results from investigating the type of responses produced by the undesirability treatment. I now examine the evidence for and against the proposition that the number of respondents who have indicated a negative desirability evaluation in the undesirability treatment is greater than what randomness would predict. In this case, the null hypothesis states that the probability of observing the given number of successes (responses within the [-3, -1] interval) is equal to 3/7 – the disjunctive probability of three equally likely, mutually exclusive and independent events. Here, our initial impressions from the frequency distributions are confirmed. Across the three Likert-type items measuring how desirable or undesirable the event in the undesirability treatment is, there is in fact not enough evidence to reject the null – in all cases the null cannot be rejected even at the 15% significance level. Therefore, the conclusion is that the undesirability treatment did not encourage responses in the hypothesized direction.

```
. bitest des1==1/7 if condition==0

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des1 |       35         17            5       0.14286      0.48571

  Pr(k >= 17) = 0.000001  (one-sided test)
  Pr(k <= 17) = 1.000000  (one-sided test)
  Pr(k >= 17) = 0.000001  (two-sided test)

  note: lower tail of two-sided p-value is empty

. bitest des2==1/7 if condition==0

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des2 |       35         19            5       0.14286      0.54286

  Pr(k >= 19) = 0.000000  (one-sided test)
  Pr(k <= 19) = 1.000000  (one-sided test)
  Pr(k >= 19) = 0.000000  (two-sided test)

  note: lower tail of two-sided p-value is empty

. bitest des3==1/7 if condition==0

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des3 |       35         20            5       0.14286      0.57143

  Pr(k >= 20) = 0.000000  (one-sided test)
  Pr(k <= 20) = 1.000000  (one-sided test)
  Pr(k >= 20) = 0.000000  (two-sided test)

  note: lower tail of two-sided p-value is empty
```

Figure 9: Binomial Test, Number of "Successful" (Un)desirability Ratings is due to Chance, Control Group

Moving on to the last group, Figure 11 shows that there is enough evidence to reject the null for the desirability treatment at the 5% significance level or below in the cases of all three Likert-type desirability items. Hence, we can conclude that the desirability treatment was consistent with the *a priori* expectations.

As a final test, I pool the two treatments to test whether they were successful in encouraging non-zero responses. Figure 12 presents the results on the probability that a randomly drawn score from the two treatments (denoted by "*cond==1*" in this case) will be lower than or greater than zero, where successes are defined as number of non-zero scores. The observed probability of producing the actual number of successes is, in fact, not much higher than the expected probability. The strongest case against the null can be observed in the Likert-type desirability items that are usually presented to the respondents last (i.e. item or question 3). However, even in that case there is almost 14% probability of observing a greater number of non-zero evaluations with the null still not being false, and hence there does is not enough evidence to reject the null hypothesis.

```
. bitest des1==3/7 if condition==1

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des1 |       32         17      13.71429      0.42857      0.53125

  Pr(k >= 17)           = 0.159815  (one-sided test)
  Pr(k <= 17)           = 0.911212  (one-sided test)
  Pr(k <= 10 or k >= 17) = 0.284485  (two-sided test)

. bitest des2==3/7 if condition==1

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des2 |       32         14      13.71429      0.42857      0.43750

  Pr(k >= 14)           = 0.527011  (one-sided test)
  Pr(k <= 14)           = 0.613291  (one-sided test)
  Pr(k <= 13 or k >= 14) = 1.000000  (two-sided test)

. bitest des3==3/7 if condition==1

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des3 |       32         14      13.71429      0.42857      0.43750

  Pr(k >= 14)           = 0.527011  (one-sided test)
  Pr(k <= 14)           = 0.613291  (one-sided test)
  Pr(k <= 13 or k >= 14) = 1.000000  (two-sided test)
```

Figure 10: Binomial Test, Number of "Successful" (Un)desirability Ratings is due to Chance, Undesirability Treatment

```
. bitest des1==3/7 if condition==2

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des1 |       35         21           15      0.42857      0.60000

  Pr(k >= 21)          = 0.030781  (one-sided test)
  Pr(k <= 21)          = 0.986415  (one-sided test)
  Pr(k <= 9 or k >= 21) = 0.058646  (two-sided test)

. bitest des2==3/7 if condition==2

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des2 |       35         26           15      0.42857      0.74286

  Pr(k >= 26)          = 0.000163  (one-sided test)
  Pr(k <= 26)          = 0.999961  (one-sided test)
  Pr(k <= 4 or k >= 26) = 0.000225  (two-sided test)

. bitest des3==3/7 if condition==2

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des3 |       35         30           15      0.42857      0.85714

  Pr(k >= 30)          = 0.000000  (one-sided test)
  Pr(k <= 30)          = 1.000000  (one-sided test)
  Pr(k <= 1 or k >= 30) = 0.000000  (two-sided test)
```

Figure 11: Binomial Test, Number of "Successful" (Un)desirability Ratings is due to Chance, Desirability Treatment

```
. bitest des1==6/7 if cond==1

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des1 |       67         58      57.42857     0.85714     0.86567

  Pr(k >= 58)           = 0.506883  (one-sided test)
  Pr(k <= 58)           = 0.631863  (one-sided test)
  Pr(k <= 57 or k >= 58) = 1.000000  (two-sided test)

. bitest des2==6/7 if cond==1

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des2 |       67         59      57.42857     0.85714     0.88060

  Pr(k >= 59)           = 0.368137  (one-sided test)
  Pr(k <= 59)           = 0.758850  (one-sided test)
  Pr(k <= 56 or k >= 59) = 0.727134  (two-sided test)

. bitest des3==6/7 if cond==1

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        des3 |       67         61      57.42857     0.85714     0.91045

  Pr(k >= 61)           = 0.139561  (one-sided test)
  Pr(k <= 61)           = 0.930387  (one-sided test)
  Pr(k <= 54 or k >= 61) = 0.292635  (two-sided test)
```

Figure 12: Binomial Test, Number of "Successful" (Un)desirability Ratings is due to Chance, Combined Treatments

5.2. TESTING THE SECONDARY HYPOTHESES

The results from the binomial test presented in the previous subsection show that the treatments were not as effective in steering responses towards a certain end of the choice spectrum as tentatively hypothesized in subsection 3.5. Nevertheless, the analysis so far does not constitute proof that there are no differences in desirability evaluations between the experimental groups. What the binomial test did tell us is that respondents allocated to the control group were significantly more likely than what random choices would predict to evaluate all events presented to them as having neutral desirability. Respondents in the desirability treatment were more likely to evaluate most stimuli as desirable, whereas respondents in the undesirability treatment were not more likely than randomness to choose a negative score.

In order to establish whether there is enough evidence to conclude that event (un)desirability evaluations differ between conditions, I use the non-parametric test of ranks called the Kruskal-Wallis H test. In Stata, this is done by using the command "kwallis *varname*, by(*groupvar*)", where "*varname"* refers to the variable recording the outcome (i.e. the ambiguity indexes) and "*groupvar*" specifies a variable that identifies the groups, i.e. the variable capturing the three experimental conditions. Because we are interested in finding out whether differences exist across the three experimental conditions in their choices between positive, negative and neutral event (un)desirability scores, the 7-category variable capturing their evaluations is collapsed or recoded into 3 buckets according to sign of scores. This is a weaker test of differences compared to using the original 7-category answer format, but it reflects the research question better since we are less interested in intensity or magnitude of score and more interested in the sign. Nevertheless, even using the arguably stricter form of the test, two of the resulting statistics are significant at the 1% significance level while the third – at the 10% significance level.

```
. recode des1_1 (-3/-1 = 0) (0 = 1) (else = 2), gen(des1)
(83 differences between des1_1 and des1)
```

Figure 13 presents the results from this test of the recoded event (un)desirability scores on the differences between the three experimental groups in the evaluation of the stimulus first presented to respondents (i.e. what I defined in subsection 3.4 as the "worst-case" scenario). Since there are ties in the data, we refer to the test statistic which takes this into account, and it gives a statistical significance level (or p-value) of approximately 0.04 which just narrowly rejects the null at the 5% significance level. Therefore, there are significant differences between conditions in the way respondents evaluated the first stimulus presented to them.

```
. kwallis des1, by(condition)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  | condit~n | Obs | Rank Sum |
  |----------+-----+----------|
  |        0 |  35 |  1864.50 |
  |        1 |  32 |  1331.50 |
  |        2 |  35 |  2057.00 |
  +---------------------------+

chi-squared =     5.815 with 2 d.f.
probability =     0.0546

chi-squared with ties =     6.647 with 2 d.f.
probability =     0.0360
```

Figure 13: Kruskal-Wallis, (Un)desirability ratings by Condition, Event 1

Similarly, Figure 14 presents the results for the other two stimuli or questions. In both cases the null hypothesis can be rejected at the 1% significance level (referring again to the chi-squared with ties test statistic), indicating that respondents evaluated the second and third stimulus presented to them differently depending on the condition they were allocated to. Finally, Figure 15 and Figure 16 present the results from an additional test which considers not only the presence of (or lack thereof) differences between conditions but also whether these differences run in the hypothesized directions. Recall that it was suggested earlier that the desirability evaluations provided by respondents allocated to the undesirability treatment would be the lowest, followed by the control group and finally the desirability treatment. This hypothesized ordering manifests itself in the numerical labeling of the conditions. Just for this test (the Jonckheere-Tepstra test for ordered alternatives), I switch around the group order to be consistent with the hypothesized hierarchy of desirability ratings across conditions. Here, group 0 is the undesirability treatment while group 1 is the neutral condition. The desirability treatment keeps its label since it is expected to produce the most positive event desirability evaluations.

```
. kwallis des2, by(condition)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  | condit~n | Obs | Rank Sum |
  |----------+-----+----------|
  |        0 |  35 |  1696.00 |
  |        1 |  32 |  1366.00 |
  |        2 |  35 |  2191.00 |
  +---------------------------+

chi-squared =     8.134 with 2 d.f.
probability =     0.0171

chi-squared with ties =     9.643 with 2 d.f.
probability =     0.0081

. kwallis des3, by(condition)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  | condit~n | Obs | Rank Sum |
  |----------+-----+----------|
  |        0 |  35 |  1604.50 |
  |        1 |  32 |  1330.00 |
  |        2 |  35 |  2318.50 |
  +---------------------------+

chi-squared =    13.578 with 2 d.f.
probability =     0.0011

chi-squared with ties =    17.097 with 2 d.f.
probability =     0.0002
```

Figure 14: Kruskal-Wallis, (Un)desirability ratings by Condition, Events 2 & 3

The results show that for all three stimuli the conditions follow the expected pattern of an ascending order which is significant at the 1% significance level. In summary, desirability evaluations do differ across conditions and in general follow the predicted pattern in that the undesirability treatment generates the lowest event desirability scores, whereas the desirability treatment – the highest, with the neutral group in-between the two.

```
. recode condition (0 = 1) (1 = 0) (else = 2)
(condition: 67 changes made)

. tab condition

  condition |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         32       31.37       31.37
          1 |         35       34.31       65.69
          2 |         35       34.31      100.00
------------+-----------------------------------
      Total |        102      100.00

. jonter des1, by(condition)

Jonckheere-Terpstra Test for Ordered Alternatives

       J  = 2133.5
       J* = 2.636 (corrected for ties)


Pr(|Z| > |J*|) = 0.0084 (ordered alternative in either direction)
    Pr(Z > J*) = 0.9958 (descending ordered alternative)
    Pr(Z < J*) = 0.0042 (ascending ordered alternative)
```

Figure 15: Jonckheere Test, (Un)desirability ratings by Condition, Event 1

```
. jonter des2, by(condition)

Jonckheere-Terpstra Test for Ordered Alternatives

       J  = 2208.5
       J* = 3.187 (corrected for ties)


Pr(|Z| > |J*|) = 0.0014 (ordered alternative in either direction)
    Pr(Z > J*) = 0.9993 (descending ordered alternative)
    Pr(Z < J*) = 0.0007 (ascending ordered alternative)

. jonter des3, by(condition)

Jonckheere-Terpstra Test for Ordered Alternatives

       J  = 2331
       J* = 4.130 (corrected for ties)


Pr(|Z| > |J*|) = 0.0000 (ordered alternative in either direction)
    Pr(Z > J*) = 1.0000 (descending ordered alternative)
    Pr(Z < J*) = 0.0000 (ascending ordered alternative)
```

Figure 16: Jonckheere Test, (Un)desirability ratings by Condition, Events 2 & 3

5.3. DESCRIPTIVE STATISTICS - AMBIGUITY INDEXES

In this part I analyze graphically the values of the two ambiguity indexes across the three conditions. The numbers in Table 1 show relatively similar mean ambiguity-aversion scores for the control group (0) and the undesirability treatment (1), both very close to ambiguity-neutrality. The desirability treatment generated a greater mean score which is thus slightly further away from ambiguity neutrality. Respondents in the desirability treatment were therefore slightly less ambiguity neutral (and more averse) compared to respondents in the other two experimental groups. Concerning the ambiguity-insensitivity index, the mean score in the control group was highest, followed by the desirability and then the undesirability treatments. Therefore, respondents in the undesirability treatment showed relatively greater discrimination (though still rather weak in absolute terms) between different levels of uncertainty whereas the mean score in the control group was in the middle between perfect discrimination and insensitivity.

Figure 17 presents density histograms (as the output from ambiguity indexes resembles a continuous variable much more than a discrete) of the a-aversion index. It is immediately noticeable that the most common attitude is in fact ambiguity neutrality or close to ambiguity neutrality. It can also be seen the relatively more individuals can be considered ambiguity averse than ambiguity seeking although the shape of the histogram appears to be quite similar across conditions.

```
. table condition, cont(mean b sd b)
```

condition	mean(b)	sd(b)
0	0.08	.3962652
1	0.06	.3189114
2	0.18	.3861866

```
. table condition, cont(mean a sd a)
```

condition	mean(a)	sd(a)
0	.5031746	.5389584
1	.37673611	.5700151
2	.45714286	.4971023

Table 1

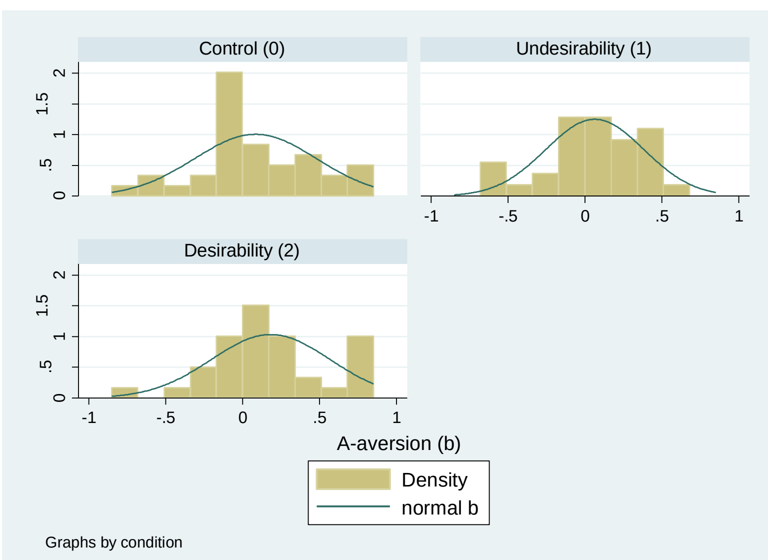


Figure 17: Histogram, A-aversion by Condition

Similarly, Figure 18 contains a density histogram of the values for the a-insensitivity index by condition. The control group and the desirability treatment appear to generate almost identical results with the most common behavior among respondents being intense insensitivity to different levels of uncertainty. The picture appears somewhat different for the undesirability treatment – here almost an equal proportion of respondents demonstrated full discrimination between the different levels of uncertainty and there were more cases of violations of the lower boundary of the index – the area beyond the lower boundary is described by $\overline{m_s} < 1/3 < 2/3 < \overline{m_c}$.

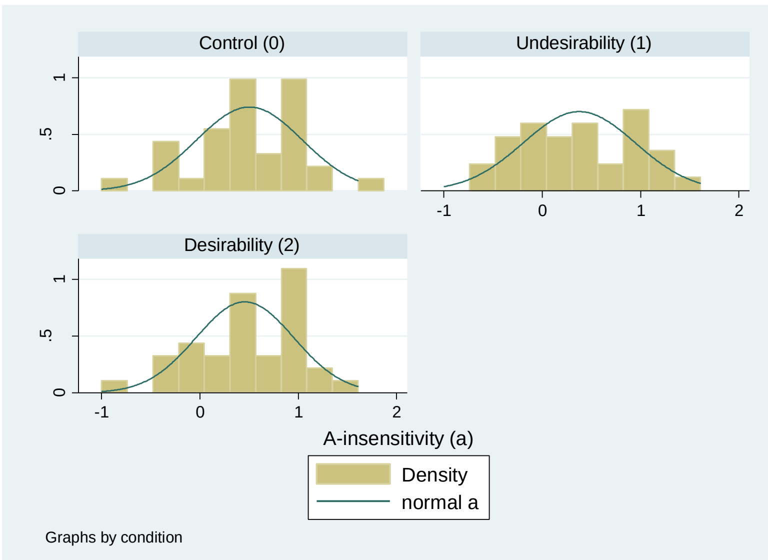


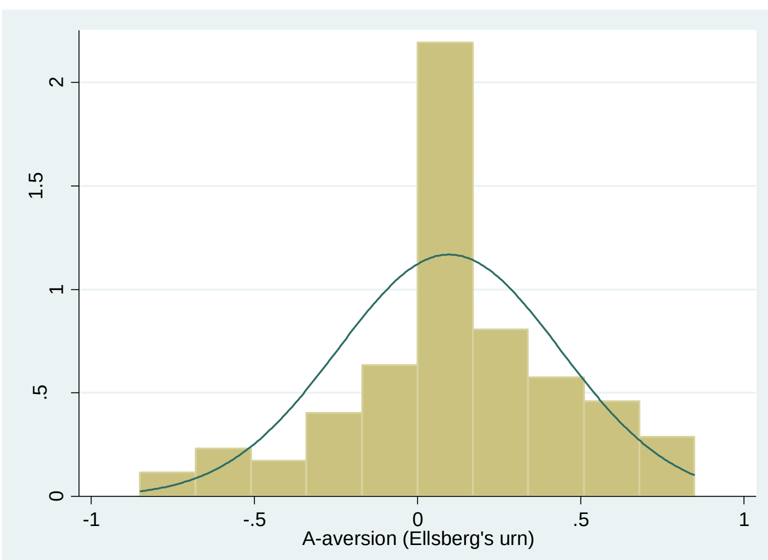
Figure 18: Histogram, A-insensitivity by Condition

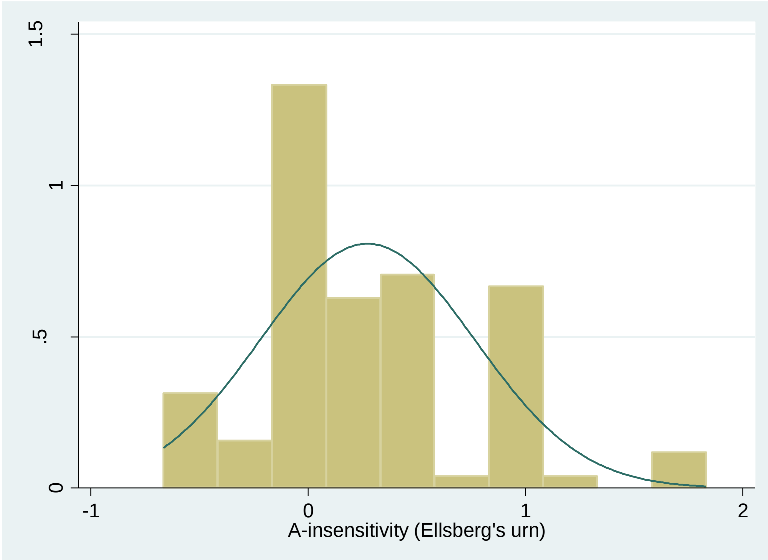
Finally, I also consider the ambiguity attitudes generated by the Ellsberg's urn problem. The numbers show that the mean ambiguity-aversion score is very similar to what the natural events in the control and undesirability group produce, and half the mean score from the desirability treatment. For the a-aversion index, respondents showed on average a significantly stronger capacity for discrimination between different levels of uncertainty compared to what any of the natural events have generated.

```
. sum b_Ellsberg, detail

                         b_Ellsberg
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.7833333           -.85
 5%    -.5166667      -.7833333
10%         -.25           -.55       Obs                 102
25%         -.05           -.55       Sum of Wgt.         102

50%            0                      Mean            .096732
                        Largest       Std. Dev.      .3415251
75%          .25            .85
90%          .55            .85       Variance       .1166394
95%          .65            .85       Skewness       .0613105
99%          .85            .85       Kurtosis       3.389858
```

```
. sum a_Ellsberg, detail

                          a_Ellsberg
-------------------------------------------------------------
      Percentiles      Smallest
 1%           -.5      -.6666667
 5%           -.5            -.5
10%     -.3333333            -.5       Obs                 102
25%     -3.70e-16            -.5       Sum of Wgt.         102

50%      .1666667                      Mean           .2734205
                        Largest        Std. Dev.      .4940476
75%            .5       1.166667
90%      .9444444       1.666667       Variance        .244083
95%      1.055556       1.666667       Skewness       .7350357
99%      1.666667       1.833333       Kurtosis       3.575996
```

The density histogram of the a-aversion index shows a clear unimodal distribution, very closely resembling the normal distribution (with a mean value close to ambiguity-neutrality) with a slightly higher proportion of respondents exhibiting ambiguity-aversion compared to ambiguity-seeking behavior. The density histogram of the a-insensitivity index also possesses a clear peak in the area of perfect discrimination. This shape is noticeably different from the ones generated by the natural events which are often bimodal and with one peak in the area of complete insensitivity to different levels of uncertainty.





5.4. TESTING THE PRIMARY HYPOTHESES

To test the main research questions of this study, I follow the statistical approaches outlined in subsection 4.3. I attempt to establish the relevance and significance of the conjectured covariate and in the case of absence of any significance impact I apply the standard nonparametric Kruskal-Wallis test. To check the robustness of these results I also perform a non-parametric (rank) analysis of covariance using the Quade test. Although of primary interest is the potential difference revealed by the between-subjects design or across the control and two treatments, I also explore the within-subjects dimension of the questionnaire, i.e. differences between ambiguity attitudes elicited by means of the Ellsberg's urn problem and a natural event. For example, such an analysis can give us a clue as to which of the three natural events resembles most the Ellsberg's urn problem. The Ellsberg's urn has traditionally been thought of as desirability-neutral event, or an aleatory event as a research psychologist would call it, and I can use this opportunity to provide empirical evidence on this topic.

Table 2 and Table 3 provide summary statistics of the ambiguity indexes per different groups of the covariate. The numbers show that greater degrees of perceived competence are associated with a lower ambiguity aversion index and greater discrimination. This appears to be in line with the established empirical findings in this area of research (e.g. Heath and Tversky, 1991).

```
. table comp, cont(mean b sd b)
```

RECODE of comp1_1 (comp1_1)	mean(b)	sd(b)
0	0.10	.371362
1	0.18	.3139331
2	-0.00	.4482154

Table 2

```
. table comp, cont(mean a sd a)
```

RECODE of comp1_1 (comp1_1)	mean(a)	sd(a)
0	.47660819	.5939624
1	.44973545	.496572
2	.39393939	.5053823

Table 3

Next, I proceed to run two ordered logistic regressions where each of the two ambiguity indexes is being predicted by the perceived competence variable. This enables us to establish whether there is a meaningful association between the two variables. Figure 19 provides evidence on this question. The likelihood ratio chi-squared of 0.44 and 0.32 for each model indicates a very high p-value and that the model is insignificant overall or, in other words, not different at all compared to the null model with no predictors. Therefore, the conclusion is that the hypothesized covariate has no explanatory power and that it has no impact on any of the two indexes describing ambiguity attitudes. This finding adds to the previous one from subsection 5.1 where it is established that perceived competence and the main regressor of interest (experimental groups) are independent and uncorrelated.

```
. ologit comp b

Iteration 0:   log likelihood = -108.53389
Iteration 1:   log likelihood =  -108.3117
Iteration 2:   log likelihood = -108.31162
Iteration 3:   log likelihood = -108.31162

Ordered logistic regression                     Number of obs     =        102
                                                LR chi2(1)        =       0.44
                                                Prob > chi2       =     0.5049
Log likelihood = -108.31162                     Pseudo R2         =     0.0020

------------------------------------------------------------------------------
        comp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           b |  -.3478242   .5210881    -0.67   0.504    -1.369138    .6734897
-------------+----------------------------------------------------------------
       /cut1 |  -.5691714   .2174332                     -.9953326   -.1430101
       /cut2 |   1.249402   .2480903                      .7631541     1.73565
------------------------------------------------------------------------------

. ologit comp a

Iteration 0:   log likelihood = -108.53389
Iteration 1:   log likelihood = -108.37437
Iteration 2:   log likelihood = -108.37434
Iteration 3:   log likelihood = -108.37434

Ordered logistic regression                     Number of obs     =        102
                                                LR chi2(1)        =       0.32
                                                Prob > chi2       =     0.5721
Log likelihood = -108.37434                     Pseudo R2         =     0.0015

------------------------------------------------------------------------------
        comp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           a |  -.1989646    .352297    -0.56   0.572     -.889454    .4915248
-------------+----------------------------------------------------------------
       /cut1 |  -.6125221   .2616572                     -1.125361   -.0996834
       /cut2 |   1.204113   .2847314                      .6460498    1.762176
------------------------------------------------------------------------------
```

Figure 19: Ordered Logistic Regression, Competence-A-aversion & Competence-A-insensitivity

Armed with these findings, I take the liberty of excluding the hypothesized covariate from the next analysis which aims to investigate whether differences exist between conditions in the ambiguity attitudes exhibited by respondents. Figure 20 presents the results from a Kruskal-Wallis on ambiguity indexes by condition (with no covariate). The resulting chi-squared test statistics generated from both tests do not allow us to reject the null hypotheses that the ranks of the ambiguity indexes do not differ significantly between the three conditions. Therefore, this offers the first piece of evidence regarding the main research question which asks whether event (un)desirability is significantly associated with ambiguity attitudes.

```
. kwallis b, by(condition)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  | condit~n | Obs | Rank Sum |
  |----------+-----+----------|
  |        0 |  35 |  1706.00 |
  |        1 |  32 |  1583.50 |
  |        2 |  35 |  1963.50 |
  +---------------------------+

chi-squared =     1.298 with 2 d.f.
probability =     0.5225

chi-squared with ties =     1.300 with 2 d.f.
probability =     0.5220

. kwallis a, by(condition)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  | condit~n | Obs | Rank Sum |
  |----------+-----+----------|
  |        0 |  35 |  1920.50 |
  |        1 |  32 |  1520.50 |
  |        2 |  35 |  1812.00 |
  +---------------------------+

chi-squared =     1.038 with 2 d.f.
probability =     0.5952

chi-squared with ties =     1.039 with 2 d.f.
probability =     0.5949
```

Figure 20: Kruskal-Wallis, A-aversion by Condition & A-insensitivity by Condition

Since the analysis in subsections 5.1 and 5.2 left a few doubts regarding the effectiveness of the conditions in deliberately steering (un)desirability scores towards certain directions depending on the experimental group, I also repeat the tests from Figure 20 with the local (un)desirability ratings as predictors instead. The local (un)desirability scores here follow the same format as the one from subsection 5.2. I also create a variable labeled "*group_rangdes*" which is a categorical variable measuring the absolute range of the within-source (un)desirability evaluations of each respondent. It was tentatively suggested earlier that the more dispersed a respondent's ratings are the less discriminating he or she may be towards different levels of uncertainty. Figure 21 and Figure 22 show two sets of Kruskal-Wallis test, each set having one of the ambiguity indexes as the explained variable.

The first figure presents the first set of three tests concerning the a-aversion index. Only the third stimulus (labeled "*des3*") has groups of observations that are statistically different from each other, at the 10% significance level, in terms of the ambiguity aversion attitudes that they provoke. Therefore, in this case, we can reject the null hypothesis that the ranks of the observations within the a-aversion index are equal across the three groups of (un)desirability ratings associated with the evaluation of the third stimulus. The second figure presents the Kruskal-Wallis test the relationship between a-insensitivity index and the (un)desirability rating dispersion. The chi-squared test statistic is again very small, and thus does not allow us to reject the null hypotheses. Therefore, the conclusion is that there is not enough evidence to disprove the hypothesis that the ranks of the observations within the a-insensitivity index are equal across the different groups of (un)desirability rating dispersion.

```
. kwallis b, by(des1)

Kruskal-Wallis equality-of-populations rank test

  +-------------------------+
  | des1 | Obs |  Rank Sum  |
  |------+-----+------------|
  |    0 |  33 |   1859.50  |
  |    1 |  26 |   1369.00  |
  |    2 |  43 |   2024.50  |
  +-------------------------+

chi-squared =     1.885 with 2 d.f.
probability =     0.3897

chi-squared with ties =     1.887 with 2 d.f.
probability =     0.3892

. kwallis b, by(des2)

Kruskal-Wallis equality-of-populations rank test

  +-------------------------+
  | des2 | Obs |  Rank Sum  |
  |------+-----+------------|
  |    0 |  24 |   1249.50  |
  |    1 |  27 |   1505.00  |
  |    2 |  51 |   2498.50  |
  +-------------------------+

chi-squared =     0.930 with 2 d.f.
probability =     0.6281

chi-squared with ties =     0.932 with 2 d.f.
probability =     0.6276

. kwallis b, by(des3)

Kruskal-Wallis equality-of-populations rank test

  +-------------------------+
  | des3 | Obs |  Rank Sum  |
  |------+-----+------------|
  |    0 |  18 |    806.50  |
  |    1 |  26 |   1646.50  |
  |    2 |  58 |   2800.00  |
  +-------------------------+

chi-squared =     5.764 with 2 d.f.
probability =     0.0560

chi-squared with ties =     5.772 with 2 d.f.
probability =     0.0558
```

Figure 21: Kruskal-Wallis, A-aversion by (Transformed) (Un)desirability Ratings, Events 1, 2 & 3

```
. recode rangdes (0/1 = 0) (2/3 = 1) (else = 2), gen(group_rangdes)
(72 differences between rangdes and group_rangdes)

. tab group_rangdes

  RECODE of |
    rangdes |
(rang(des)) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         50       49.02       49.02
          1 |         34       33.33       82.35
          2 |         18       17.65      100.00
------------+-----------------------------------
      Total |        102      100.00
```

```
. kwallis a, by(group_rangdes)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  | group_~s | Obs | Rank Sum |
  |----------+-----+----------|
  |        0 |  50 |  2643.50 |
  |        1 |  34 |  1603.00 |
  |        2 |  18 |  1006.50 |
  +---------------------------+

chi-squared =     1.244 with 2 d.f.
probability =     0.5368

chi-squared with ties =     1.245 with 2 d.f.
probability =     0.5365
```

Figure 22: Kruskal-Wallis, A-insensitivity by (Transformed) (Un)desirability Ratings Dispersion (Within-Source)

To test the robustness of these result, I also proceed to perform the Quade test or to conduct a rank analysis of the covariance between each of the response variables and the experimental group, also taking into account the degree of competence. Step one involves ranking the dependent variable(s) and the covariate, ignoring the condition variable. I use a “minus” sign in front of the variable to instruct Stata to assign the highest rank (i.e. rank 1) to the biggest value (i.e. maximum aversion and maximum competence receive 1’s).

```
. egen rank_comp1_1 = rank(-comp1_1)
```

```
. egen rank_b = rank(-b)

. egen rank_a = rank(-a)
```

Step two requires regressing the ranks of the response variable(s) on the ranks of the covariate and saving the residuals from this equation, again ignoring the condition variable.

```
. reg rank_b rank_comp1_1

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(1, 100)       =      0.03
       Model |  29.3417353         1  29.3417353   Prob > F        =    0.8557
    Residual |  88266.6583       100  882.666583   R-squared       =    0.0003
-------------+----------------------------------   Adj R-squared   =   -0.0097
       Total |       88296       101  874.217822   Root MSE        =     29.71

------------------------------------------------------------------------------
      rank_b |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rank_comp1_1 |    .018518   .1015662     0.18   0.856    -.1829864    .2200223
       _cons |   50.54632   6.001114     8.42   0.000     38.64029    62.45236
------------------------------------------------------------------------------

. predict rb, residuals
```

```
. reg rank_a rank_comp1_1

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(1, 100)       =      0.12
       Model |  110.238081         1  110.238081   Prob > F        =    0.7245
    Residual |  88227.7619       100  882.277619   R-squared       =    0.0012
-------------+----------------------------------   Adj R-squared   =   -0.0087
       Total |       88338       101  874.633663   Root MSE        =    29.703

------------------------------------------------------------------------------
      rank_a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rank_comp1_1 |  -.0358936   .1015438    -0.35   0.724    -.2373535    .1655664
       _cons |   53.34852   5.999791     8.89   0.000      41.4451    65.25193
------------------------------------------------------------------------------

. predict ra, residuals
```

In the final step, I run a regular one-way ANOVA using the residuals from step two as dependent and the experimental groups as the factor variable:

```
. anova rb condition

                         Number of obs =        102    R-squared     =  0.0134
                         Root MSE      =    29.6586    Adj R-squared = -0.0065

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  1182.9868          2   591.49342      0.67  0.5128
                         |
               condition |  1182.9868          2   591.49342      0.67  0.5128
                         |
                Residual |   87083.67         99   879.63303
              -----------+----------------------------------------------------
                   Total |  88266.657        101    873.9273

. anova ra condition

                         Number of obs =        102    R-squared     =  0.0077
                         Root MSE      =    29.7371    Adj R-squared = -0.0123

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  682.70662          2   341.35331      0.39  0.6808
                         |
               condition |  682.70662          2   341.35331      0.39  0.6808
                         |
                Residual |  87545.055         99   884.29349
              -----------+----------------------------------------------------
                   Total |  88227.762        101    873.5422
```

The F-statistics in both models are very high implying that neither model is significant at any of the most common statistical significance levels. These results are consistent with the first approach followed at the start of this subsection.

Finally, Figure 23 and Figure 24 present the results from Wilcoxon signed-rank tests on the differences between a-aversion indexes produced by the Ellsberg's urn problem and by the natural source of uncertainty problems (per condition). The outcomes from these tests should help us clarify how respondents perceive the Ellsberg's urn problem, i.e. does it resemble sources containing predominantly neutral or desirable events, for example. Unfortunately, in this case, the tests show that the ambiguity attitudes that respondents demonstrated when answering questions related to the Ellsberg's urn problem are not significantly different from the ambiguity attitudes generated by any of the three natural event conditions. In neither case can we reject the null hypothesis that the a-aversion index from the Ellsberg's urn problem is identical to the a-aversion index from the natural event. This raises some interesting questions which I discuss in subsection 5.7.

```
. signrank b_Ellsberg = b if condition == 0

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       13       261.5         297
    negative |       14       332.5         297
        zero |        8          36          36
-------------+---------------------------------
         all |       35         630         630

unadjusted variance     3727.50
adjustment for ties       -0.63
adjustment for zeros     -51.00
                     ----------
adjusted variance       3675.88

Ho: b_Ellsberg = b
             z =  -0.586
    Prob > |z| =   0.5582
```

Figure 23: Wilcoxon Signed-Rank, Ellsberg's Urn A-aversion Index is Equal to Natural Event A-Aversion index, Control Group

```
. signrank b_Ellsberg = b if condition == 1

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       17       285.5       263.5
    negative |       14       241.5       263.5
        zero |        1           1           1
-------------+---------------------------------
         all |       32         528         528

unadjusted variance     2860.00
adjustment for ties       -0.88
adjustment for zeros      -0.25
                     ----------
adjusted variance       2858.88

Ho: b_Ellsberg = b
             z =   0.411
    Prob > |z| =   0.6807

. signrank b_Ellsberg = b if condition == 2

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       14         269         301
    negative |       14         333         301
        zero |        7          28          28
-------------+---------------------------------
         all |       35         630         630

unadjusted variance     3727.50
adjustment for ties        0.00
adjustment for zeros     -35.00
                     ----------
adjusted variance       3692.50

Ho: b_Ellsberg = b
             z =  -0.527
    Prob > |z| =   0.5985
```

Figure 24: Wilcoxon Signed-Rank, Ellsberg's Urn A-aversion Index is Equal to Natural Event A-Aversion index, Treatments

Similarly, Figure 25 shows that the null hypothesis of equal ranks cannot be rejected also when giving no consideration to the grouping variable "condition". The resulting z-score for equal ranks of the a-aversion indexes from the Ellsberg and the natural event problems is very low, and hence no within-subject difference appears to exist in responses across all conditions.

```
. signrank b = b_Ellsberg

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       42        2673      2558.5
    negative |       44        2444      2558.5
        zero |       16         136         136
-------------+---------------------------------
         all |      102        5253        5253

unadjusted variance    89738.75
adjustment for ties       -8.75
adjustment for zeros    -374.00
                     ----------
adjusted variance      89356.00

Ho: b = b_Ellsberg
             z =   0.383
    Prob > |z| =   0.7017
```

Figure 25: Wilcoxon Signed-Rank, Ellsberg's Urn A-aversion Index is Equal to Natural Event A-Aversion index, Pooled

5.5. A DIGRESSION

Parallel to the main approach, I also consider out of curiosity the results from a regular multiple regression analysis on the raw (untransformed) data and use a summated desirability score. Therefore, it should be taken as a minor digression and requires that we temporarily relax the conclusions reached in subsection 4.1.2 (i.e. treating data derived from individual Likert-type items as ordinal and refraining from creating summated scores) that encouraged us to take the nonparametric route to testing the tentative hypotheses and using single-event or local desirability scores.

Figure 26 and Figure 27 present the results from both regression analyses. The first thing to notice is the extremely low value for the R-squared and the overall statistical insignificance of the a-insensitivity model. This implies that the current specification and regressors explain a negligibly small fraction of the total variation in the dependent variable and that the model may even be incorrectly specified. The a-aversion model does not suffer from these issues. The two treatments in the a-insensitivity model produce ambiguity aversion coefficients compared of contrasting sign as compared to the neutral condition. This means that respondents in the undesirability treatment discriminated more whereas respondents in the desirability treatment discriminated less between different levels of uncertainty compared to respondents in the control group. However, the coefficients are very small and insignificant. The coefficients of the treatments also showed contrasting signs compared to the base group in the a-aversion model with respondents in the desirability group having a higher a-aversion index on average compared the control group, holding everything else fixed. However, the coefficients are again not significant.

It is also worth pointing out that individuals with higher degrees of perceived competence showed less ambiguity-aversion compared to respondents who indicated lower degrees of competence. This can be seen from the negative coefficient on the "2.competence" value of the variable which includes answer options 2 and 3 (highest degrees of perceived competence) from the perceived competence question. From the seminal study of Heath and Tversky (1991) on the impact of the perceived level of competence on ambiguity attitudes, we know that the more knowledgeable an individual considers herself in a topic the more ambiguity-seeking she tends to be. The results from the present study (including the non-parametric tests performed earlier in this sub-section), although lacking statistical significance, appear consistent with previous findings. This can be seen more clearly if competence is recoded into three buckets according to sign. However, this has the drawback that the bucket containing only neutral responses ("1.competence") has very few observations which is why this format is not preferred.

Finally, encouraged by the results from the Spearman rank-correlation test, I proceed to include a summated (un)desirability score (a simple average of the three local scores) instead of the local scores in the main regression. The coefficient turns out to be significant at the 5% significance level in the aversion model but not in the insensitivity model. In both models the coefficient is small and negative. This means that an increase of one division in the (un)desirability answer spectrum is associated with a lower aversion and higher ambiguity discrimination on average, holding all else fixed. Also, the variable measuring within-source (un)desirability evaluation dispersion is not significant in either model. The ambiguity indexes from the Ellsberg's urn problem turn out to be the most powerful and consistently significant at the 1% significance level in both models.

```
. spearman condition globdes

 Number of obs =     102
Spearman's rho =       0.2947

Test of Ho: condition and globdes are independent
    Prob > |t| =       0.0026
```

```
. spearman condition group_rangdes

 Number of obs =     102
Spearman's rho =       0.2184

Test of Ho: condition and group_rangdes are independent
    Prob > |t| =       0.0274
```

```
. reg b b_Ellsberg i.condition globdes i.group_rangdes i.competence

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(8, 93)        =      7.96
       Model |  5.62272278         8  .702840348   Prob > F        =    0.0000
    Residual |  8.21301796        93  .088312021   R-squared       =    0.4064
-------------+----------------------------------   Adj R-squared   =    0.3553
       Total |  13.8357407       101  .136987532   Root MSE        =    .29717

------------------------------------------------------------------------------
           b |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  b_Ellsberg |   .6107165   .0887482     6.88   0.000     .4344802    .7869527
             |
   condition |
          1  |  -.1140507   .0806602    -1.41   0.161    -.2742259    .0461245
          2  |   .0687327   .0802555     0.86   0.394    -.0906387    .2281042
             |
     globdes |  -.0627721   .0294872    -2.13   0.036    -.1213279   -.0042164
             |
group_rangdes |
          1  |   .0338397   .0729854     0.46   0.644    -.1110948    .1787742
          2  |  -.0348782   .0875761    -0.40   0.691     -.208787    .1390306
             |
  competence |
          1  |   .0856053   .0675694     1.27   0.208    -.0485743    .2197848
          2  |  -.0735654   .0809845    -0.91   0.366    -.2343846    .0872538
             |
       _cons |    .073982    .064238     1.15   0.252     -.053582     .201546
------------------------------------------------------------------------------
```

Figure 26: OLS, A-aversion, Model 1

```
  competence |
          1  |  -.0505524   .1475177    -0.34   0.733    -.3434934    .2423885
          2  |  -.0060874   .0652305    -0.09   0.926    -.1356222    .1234474
```

```
. reg a a_Ellsberg i.condition globdes i.group_rangdes i.competence

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(8, 93)        =      1.37
       Model |  3.00958863         8  .376198578   Prob > F        =    0.2218
    Residual |  25.6127788        93  .275406224   R-squared       =    0.1051
-------------+----------------------------------   Adj R-squared   =    0.0282
       Total |  28.6223675       101  .283389777   Root MSE        =    .52479

------------------------------------------------------------------------------
           a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  a_Ellsberg |   .3161147   .1091232     2.90   0.005     .0994177    .5328116
             |
   condition |
          1  |   -.029132   .1443424    -0.20   0.840    -.3157675    .2575034
          2  |   .0007345     .14078     0.01   0.996    -.2788267    .2802957
             |
     globdes |  -.0170048   .0520143    -0.33   0.744    -.1202949    .0862852
             |
group_rangdes |
          1  |  -.1561912    .130317    -1.20   0.234    -.4149748    .1025925
          2  |   .0402096   .1547526     0.26   0.796    -.2670984    .3475176
             |
  competence |
          1  |   .0104043   .1193644     0.09   0.931    -.2266297    .2474383
          2  |  -.0580502   .1436063    -0.40   0.687    -.3432238    .2271234
             |
       _cons |   .4326525   .1183407     3.66   0.000     .1976513    .6676536
------------------------------------------------------------------------------
```

Figure 27: OLS, A-insensitivity, Model 1

Finally, before I close this digression, I perform several basic tests to diagnose whether key assumption of the MLR method have been violated or not, and thus attempt to provide a degree of confidence regarding the reliability of the results from this analysis. I always report the results in the order in which the two regression are presented above, i.e. results from tests associated with the a-aversion regression appear first.

First, I inspect the variance inflation factors of the regressors to detect potential issues with multicollinearity. The resulting scores are sufficiently low so that we can be reasonably assured that the standard errors in these regressions are not inflated. The highest values circle around 1.50 and it is not a surprise that this concerns the variables capturing the experimental conditions and the summated (un)desirability score.

. vif

Variable	VIF	1/VIF
b_Ellsberg	1.05	0.951775
condition		
1	1.62	0.618092
2	1.68	0.596387
globdes	1.35	0.741018
group_rang~s		
1	1.37	0.731410
2	1.29	0.776778
competence		
1	1.28	0.782923
2	1.28	0.780375
Mean VIF	1.36	

. vif

Variable	VIF	1/VIF
a_Ellsberg	1.07	0.938169
condition		
1	1.66	0.601919
2	1.65	0.604434
globdes	1.35	0.742684
group_rang~s		
1	1.40	0.715458
2	1.29	0.775794
competence		
1	1.28	0.782391
2	1.29	0.773953
Mean VIF	1.37	

Next, I turn my attention to the residuals from the two regressions. Figure 28 and Figure 29 present the kernel density estimates of the residuals. In both cases, the deviations from the standard normal distribution are minimal, and thus it is reasonable to assume that this does not constitute a violation of normality.

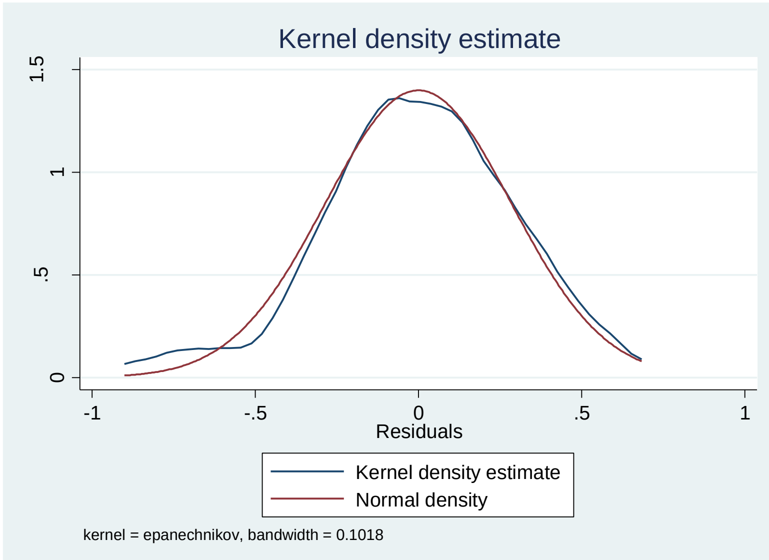


Figure 28: Kernel Density Plot, A-aversion, Model 1

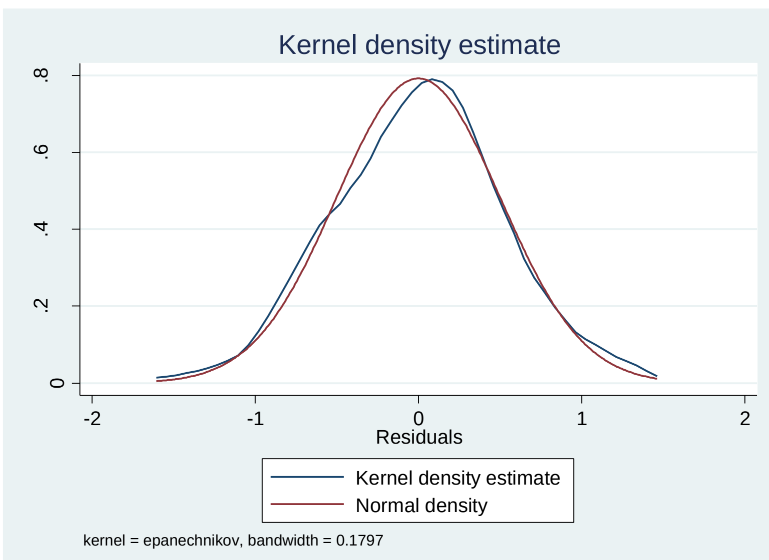


Figure 29 Kernel Density Plot, A-insensitivity, Model 1

I also plot the cumulative distribution and the quantiles of the residuals against the theoretical standard of the normal distribution. These two visual methods enable us to eyeball with greater precision any deviations in the center and the tails of the data. The kernel densities examined earlier indicate slight deviations from normality in the region of the lower tail for the residuals from the a-aversion regression model and slight deviations in the center of the distribution of the residuals from the a-insensitivity regression. The P-P plots (Figure 30 and Figure 31) and Q-Q plots (Figure 32 and Figure 33) confirm this – there is indeed a slight deviation from normality in the second plot around the central area and a more visible deviation from normality in the lower-tail area of the third plot. Nevertheless, it is safe to regard these as trivial.

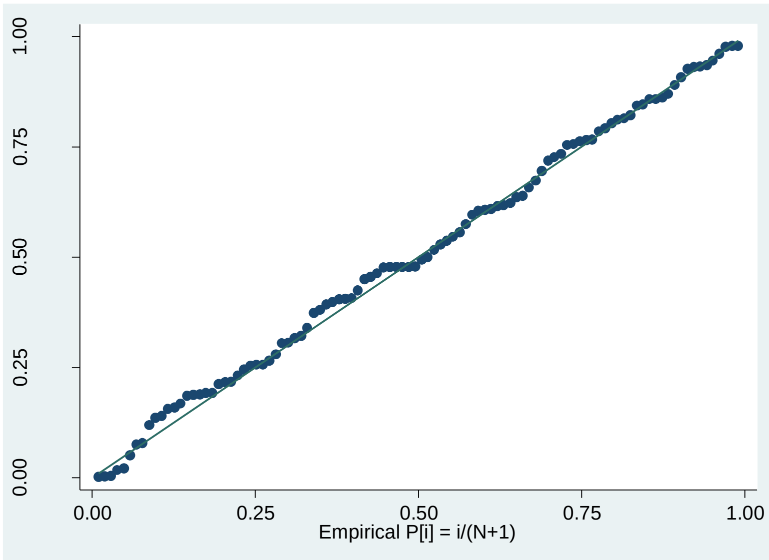


Figure 30: P-P Plot, A-aversion, Model 1

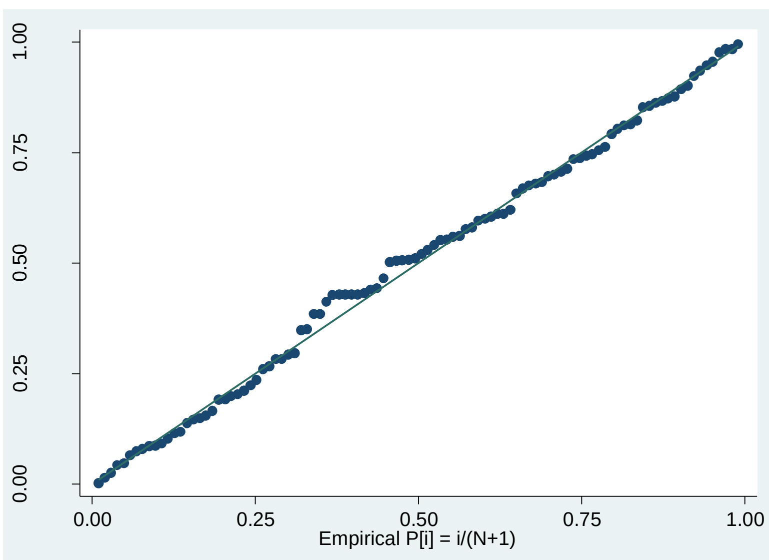


Figure 31: P-P Plot, A-insensitivity, Model 1

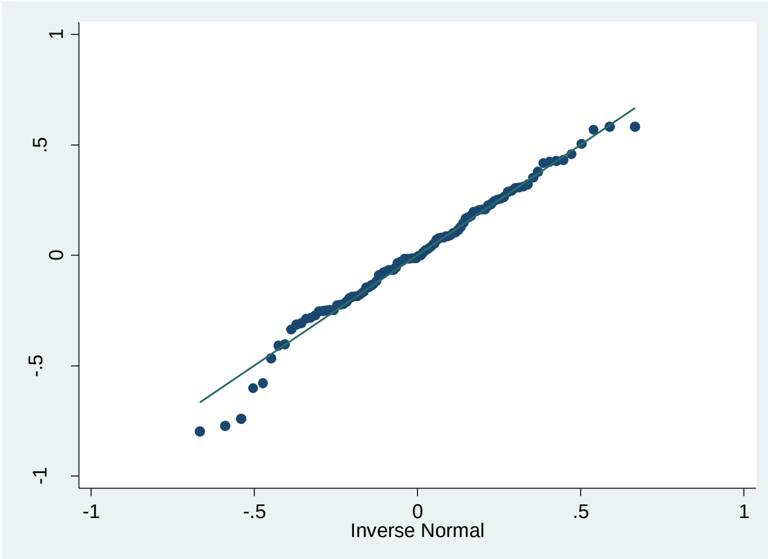


Figure 32: Q-Q Plot, A-aversion, Model 1

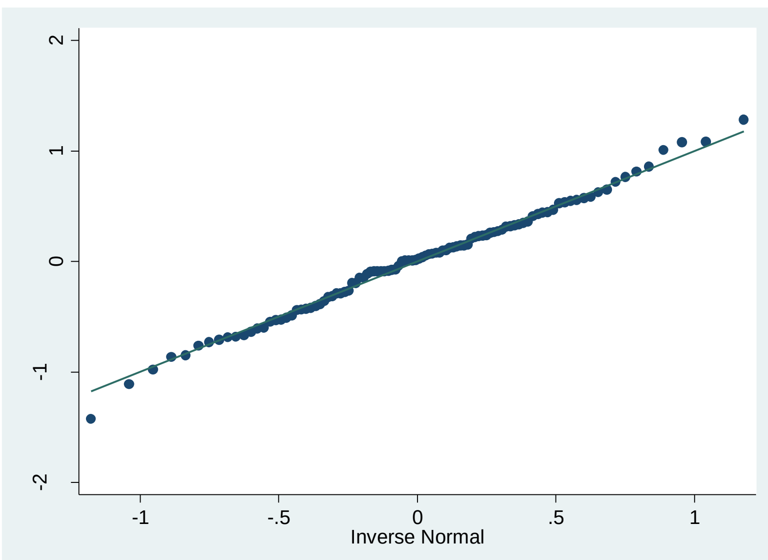


Figure 33: Q-Q Plot, A-insensitivity, Model 1

In addition to these visual assessments of normality, I also perform a numerical test following the Shapiro-Wilk method. The null hypothesis of this test is that the data are normally distributed, and therefore a very low p-value would constitute strong evidence against the assumption of normally distributed residuals. In both cases, we can reject the null at the most common significance levels.

```
. swilk resid

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
       resid |        102    0.98179      1.529     0.942    0.17300
```

```
. swilk resid

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
       resid |        102    0.99579      0.353    -2.310    0.98955
```

Next, I turn to the assumption of variance homogeneity and inspect scatter plots of the residuals as well as results from numerical tests for homoscedasticity. Figure 34 and Figure 35 plot the regression residuals against the regression fitted values. If the variance of the residuals is constant, we should observe no patterns in the plots and the observations should appear to randomly cluster around the reference line at the value 0. In both cases, it appears that the plot is slightly wider at the center than at the ends. Such shapes may be problematic as that may be indicative of heteroscedasticity or non-constant variance.

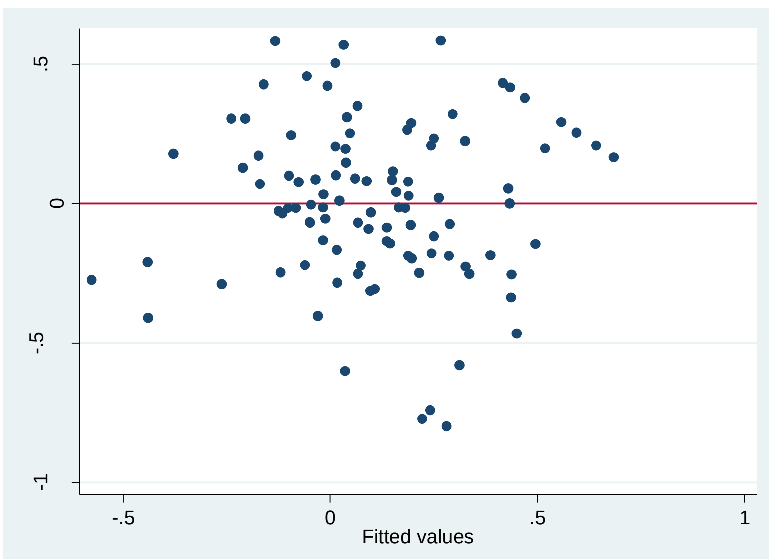


Figure 34: Residual Variance Plot, A-aversion, Model 1

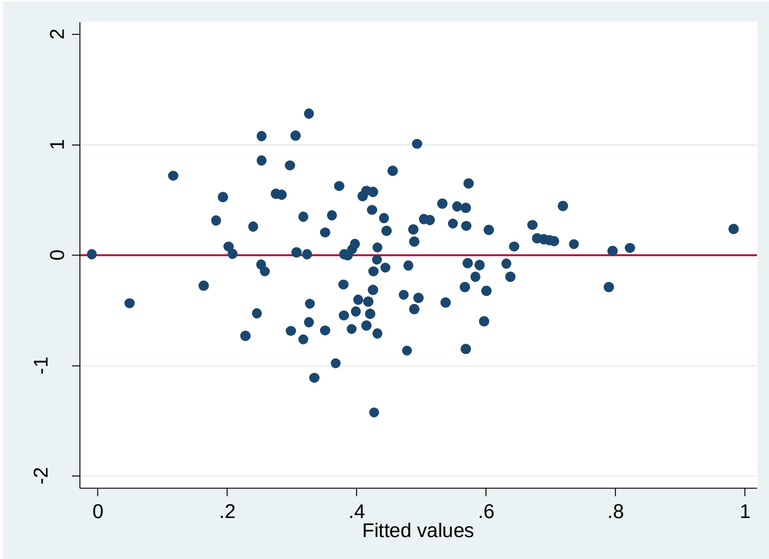


Figure 35: Residual Variance Plot, A-insensitivity, Model 1

To obtain clearer evidence regarding the satisfaction or violation of this assumption, I perform two numerical tests – the White and the Breusch-Pagan tests. These two tests should be able to identify the most common forms of heteroscedasticity, if such violation exists in the data. Both tests use the null hypothesis that the variance of the residuals is homogenous. Hence, very low p-values would be evidence for non-constant variance or some form of heteroscedasticity. In three out of four instances (two tests x two regressions) we cannot reject the null at the 5% or even at the 10% significance level and conclude that there is no evidence that the variance of residuals from each regression is not homogenous. However, the a-insensitivity model fails one of the two tests. The p-value from the Breusch-Pagan test is below the critical 1% significance level which means that we can reject the null hypothesis of homogenous residual variance. Nevertheless, the results from this test can be misleading since there are very few observations at the extreme ends as we can see from Figure 35. In this case, combining numerical and visual methods, we can conclude that the data do not appear to show too strong an evidence for heteroskedasticity in order to justify going into more detail to correct for it.

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(35)      =     33.17
         Prob > chi2   =    0.5566

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      33.17     35    0.5566
            Skewness |      18.94      8    0.0152
            Kurtosis |       0.77      1    0.3814
---------------------+-----------------------------
               Total |      52.88     44    0.1687
---------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of b

         chi2(1)      =     0.85
         Prob > chi2  =   0.3559
```

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(35)      =     44.88
         Prob > chi2   =    0.1224

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      44.88     35    0.1224
            Skewness |       5.96      8    0.6515
            Kurtosis |       0.01      1    0.9154
---------------------+-----------------------------
               Total |      50.85     44    0.2218
---------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of a

         chi2(1)      =     6.68
         Prob > chi2  =   0.0097
```

Sometimes with econometric models what may appear to be an issue with the homoscedasticity assumption would, in fact, turn out to be specification issues. Therefore, I investigate whether functional form misspecification may be a problem or perhaps even omitted variable bias by running the link test and RESET test commands. The former is built on the assumption that if the model is correctly specified, then no additional variables should be significant (except by chance). It tests this by creating an extra variable labeled “*_hatsq*”, which, if the original model is correctly specified, should have a very large p-value, and hence fail the t-test. This is exactly what we see in both models – the p-values are 0.86 and 0.33 for the a-aversion and a-insensitivity model, respectively. The Ramsey RESET test, in contrast, is widely considered a test for omitted variable bias instead. Indeed, Stata generates a null hypothesis for this test which refers to the model having no omitted variables. The null cannot be rejected for the a-insensitivity model since the p-value is very large (~0.50). However, in the case of the a-aversion model, we reject the null at the 5% significance level, implying that there are potentially important variables which are lacking from the original model.

```
. linktest

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(2, 99)        =     33.92
       Model |   5.6256396         2   2.8128198   Prob > F        =    0.0000
    Residual |  8.21010115        99  .082930315   R-squared       =    0.4066
-------------+----------------------------------   Adj R-squared   =    0.3946
       Total |  13.8357407       101  .136987532   Root MSE        =    .28798

------------------------------------------------------------------------------
           b |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .9865771   .1409673     7.00   0.000     .7068674    1.266287
      _hatsq |   .0645098   .3439752     0.19   0.852    -.6180116    .7470312
       _cons |  -.0028611   .0350409    -0.08   0.935    -.0723898    .0666677
------------------------------------------------------------------------------

. ovtest

Ramsey RESET test using powers of the fitted values of b
       Ho:  model has no omitted variables
                  F(3, 90) =      3.24
                  Prob > F =      0.0257
```

```
. linktest

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(2, 99)        =      6.35
       Model |  3.25602091         2  1.62801045   Prob > F        =    0.0025
    Residual |  25.3663466        99  .256225723   R-squared       =    0.1138
-------------+----------------------------------   Adj R-squared   =    0.0959
       Total |  28.6223675       101  .283389777   Root MSE        =    .50619

------------------------------------------------------------------------------
           a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |  -.0320739   1.092082    -0.03   0.977    -2.199002    2.134855
      _hatsq |   1.110456   1.132306     0.98   0.329    -1.136285    3.357196
       _cons |   .2067206       .253     0.82   0.416    -.2952863    .7087275
------------------------------------------------------------------------------

. ovtest

Ramsey RESET test using powers of the fitted values of a
       Ho:  model has no omitted variables
                  F(3, 90) =      0.81
                  Prob > F =      0.4922
```

To conclude, the tests show that there should not be very serious concerns about violations of major assumptions. However, the results from the Ramsey RESET test on the a-aversion model may suggest omitted variables. Including interactions or non-linear transformations of independent variables did not improve the specification. Unfortunately, since the present study uses primary data, additional variables containing new information are not readily available. Nevertheless, I do attempt adding variables such as respondent demographics and other information extracted from respondents' answers.

Figure 36 presents the new and expanded regression analysis. The new variable "*is_bef*" is a binary variable indicating whether the respondent holds or currently pursues a degree in Business, Economics or Finance (affirmation is associated with the value "1"). The new variable "*Time*" captures the number of minutes which the respondent took to complete the questionnaire, while "*group_rangdes*" was introduced earlier as a simple measure of the within-source (un)desirability ratings dispersion, taken as the highest (i.e. most positive) minus the lowest local (un)desirability score. Adding these extra variables makes the undesirability treatment indicator significant at the 10% level whereas the statistical significance of the global (un)desirability index improves further. The coefficients and standard errors do not experience large changes though. Interestingly, although not shown here, the summated (un)desirability score loses statistical significance even at the 10% significance level when the predictor "condition" is excluded from the regression (the inclusion or exclusion of any other variable has only trivial impact on its coefficient and standard error). Using the local (un)desirability evaluations instead does not produce significant association with the ambiguity indexes. Also, using the alternative format of the competence variable (the one with three buckets according to sign rather than strength or intensity) results merely in trivial differences.

```
. reg b b_Ellsberg i.condition globdes i.group_rangdes i.competence a_Ellsberg a Age i.Female i.
> is_bef Time

      Source |       SS           df       MS      Number of obs   =       102
-------------+----------------------------------   F(14, 87)       =      5.15
       Model |  6.27116341        14  .447940244   Prob > F        =    0.0000
    Residual |  7.56457733        87  .086949165   R-squared       =    0.4533
-------------+----------------------------------   Adj R-squared   =    0.3653
       Total |  13.8357407       101  .136987532   Root MSE        =    .29487

-------------------------------------------------------------------------------
            b |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
   b_Ellsberg |   .5786528   .0933549     6.20   0.000     .3930999    .7642058
              |
    condition |
           1  |  -.1489772   .0830648    -1.79   0.076    -.3140774     .016123
           2  |   .0981598   .0816279     1.20   0.232    -.0640845    .2604041
              |
      globdes |  -.0780139   .0306145    -2.55   0.013    -.1388637   -.0171642
              |
group_rangdes |
           1  |   .0637853   .0748525     0.85   0.396    -.0849922    .2125628
           2  |  -.0409089   .0879912    -0.46   0.643    -.2158009    .1339832
              |
   competence |
           1  |   .1012434   .0711241     1.42   0.158    -.0401235    .2426104
           2  |   -.056449   .0843569    -0.67   0.505    -.2241175    .1112195
              |
   a_Ellsberg |  -.0446198   .0645233    -0.69   0.491    -.1728669    .0836273
            a |   .0650409   .0625361     1.04   0.301    -.0592564    .1893383
          Age |  -.0026083   .0044412    -0.59   0.559    -.0114358    .0062191
     1.Female |   .1193246   .0684163     1.74   0.085    -.0166601    .2553094
     1.is_bef |   .1510052   .0747929     2.02   0.047     .0023462    .2996643
         Time |   -.004163   .0121603    -0.34   0.733     -.028333     .020007
        _cons |   .0381907   .1538753     0.25   0.805    -.2676532    .3440346
-------------------------------------------------------------------------------
```

Figure 36: OLS, A-aversion, Model 2

Similarly, this model does not show any meaningful deviations from the key assumptions of the linear regression method. Moreover, the Ramsey RESET test has an insignificant F-statistic this time at the 5% or even the 10% significance level.

```
. ovtest

Ramsey RESET test using powers of the fitted values of b
       Ho:  model has no omitted variables
                  F(3, 84) =      1.81
                  Prob > F =      0.1516
```

Because of the linear association between the ambiguity indexes and the associated matching probabilities, using the individual or average scores for the single and composite events as dependent variables in place of the a-aversion and a-insensitivity indexes does not change the regression results in any way. Including more variables in the a-insensitivity model does not lead to any meaningful changes in coefficient size or statistical significance.

5.6. SUMMARY OF KEY RESULTS

Examining the effectiveness of the different experimental group by looking at the frequency distribution of event (un)desirability scores per condition and question reveals the following information. In the undesirability treatment, 17 respondents (53%) provided a negative event evaluation on the first question, 14 on the second and the third (44%). In the desirability treatment, 21 (60%) gave a positive score on the first question, 26 (74%) on the second, and 30 (86%) on the third. Finally, in the control group, 17 (49%) respondents gave a neutral score on the first event (un)desirability rating question, 19 (54%) on the second, and 20 (57%) on the third. Since all respondents provide three event (un)desirability ratings, there is respondent overlap in the quoted numbers as mentioned at the beginning of subsection 5.1.

Results from the binomial test indicated that respondents allocated to the control group were significantly more likely than what random choices would predict to evaluate all events presented to them as having neutral desirability, respondents in the desirability treatment were more likely to evaluate a stimulus as desirable, whereas respondents in the undesirability treatment were not more likely than randomness to choose a negative score. Finally, results from the Kruskal-Wallis and Jonckheere-Tepstra tests show that these differences between the experimental groups in the sign of event (un)desirability ratings are statistically significant and ranked according to our prior expectations.

Turning to the ambiguity indexes, the mean a-version index is 0.08 for the control group, 0.06 for the undesirability treatment, and 0.18 for the desirability treatment. The density graphs per condition reveal that ambiguity-neutrality is by far the most common attitude. For the a-insensitivity index, the mean score is 0.5 in the control group, 0.38 in the undesirability treatment, and 0.48 in the desirability treatment. Density graphs of the scores also reveal that the most common attitude among respondents is insensitivity to different levels of uncertainty for the control and desirability groups, whereas for the undesirability treatment almost an equal proportion of respondents demonstrated full discrimination. Interestingly, the graphs for both indexes often have a bimodal shape suggesting two very different groups of respondents even within conditions.

However, output from a Kruskal-Wallis test on the differences in the a-aversion and a-insensitivity indexes between conditions suggests that we cannot reject the null hypothesis. Hence, differences in ambiguity attitudes between groups are not significant. I also tested for differences in ambiguity attitudes based on the local event (un)desirability ratings and their dispersion rather than condition allocation. There is a significant difference in the a-aversion index depending on the sign of the (un)desirability score only for the third stimulus, whereas there is not enough evidence to disprove the null hypothesis of equal ranks across different degrees of (un)desirability ratings dispersion in the case of the a-insensitivity index.

Further, the results from the peripheral parametric analysis show that the experimental group categorical independent variable has no effect on any of the two ambiguity indexes. The summated (un)desirability score is negatively and significantly associated with the a-aversion index but not with the a-insensitivity index. Hence, an increase in the reported event (un)desirability is associated with a decrease in the a-aversion index, all else fixed, whereas it has no impact on a-insensitivity. Interestingly, the summated (un)desirability score loses its significance if the grouping regressor is omitted from the analysis, implying a negative confounding effect. The dispersion of (un)desirability scores is insignificant across all models

Finally, the perceived competence categorical variable turns out to have no significant association with any of the ambiguity indexes and main independent variables of interest. Results from ordered logistic regressions of perceived competence on a-aversion or a-insensitivity produce very large p-values for both models, whereas the Spearman correlation coefficient and Chi-Squared test of independence show that the conjectured covariate is uncorrelated with and independent from the experimental group categorical variable.

5.7. DISCUSSION OF KEY RESULTS

In summary, the treatments were not fully effective at generating event (un)desirability scores of the predicted sign. Most noticeably, the undesirability treatment provoked an unexpectedly large proportion of positive evaluations which effectively prohibited this research from studying ambiguity attitudes generated by undesirable events. Nevertheless, the experiment was effective at inducing 80% of the respondents to give all three mutually exclusive and exhaustive events a desirability rating of consistent or at least of non-contrasting sign. This was precisely the behavior that I wished to observe and which I hypothesized may be an important clue in understanding how and why, if at all, does desirability bias impact ambiguity attitudes. However, perhaps partly due to this ineffectiveness of the different experimental conditions in generating groups of responses which are sufficiently different from each other, I was not able to establish a clear relationship between the exogenous allocation to a given condition and certain consequences or changes in respondents' attitudes towards ambiguity.

The only significant association (I distance myself from any claims about causation) between the two comes from a regular multiple regression analysis which arguably stands on a shaky theoretical foundation. The negative coefficient on the summated (un)desirability score appears to contradict the "superstitious beliefs" hypothesis which predicts seemingly pessimistic choices in response to desirable events. Instead, the results appear more consistent with what gets loosely referred to by laymen as the "wishful thinking" effect.

An important consequence of the ineffectiveness of the conditions to serve their intended purpose is the resulting confounding influence on the event (un)desirability scores. It appears contradictory that the coefficient representing the undesirability treatment's impact on ambiguity aversion relative to the neutral condition is negative while that of the desirability treatment is positive. It appears contradictory precisely because the coefficient of the summated (un)desirability score shows that increases in desirability are associated with lower a-aversion, holding everything else (including condition) fixed. To elaborate, the highest or most positive event (un)desirability evaluations are generated by the desirability treatment followed by the control and then the undesirability group. Therefore, based on these facts, the signs on the grouping variable seem puzzling. If the desirability treatment contains the most positive event (un)desirability evaluations and higher scores are associated with lower a-aversion then it should follow that the desirability treatment has lower a-aversion compared to the control group, and hence a negative sign before the regression coefficient (however mean "*b*" is highest in the desirability treatment). Nevertheless, perhaps it is futile to go into greater depths in attempting to explain this seeming compensating confound effect given the lack of robust evidence regarding the statistical insignificance of the variable.

A further interesting observation is the consistent significance of the ambiguity attitude indexes from the Ellsberg's urn problem in predicting the ambiguity attitudes generated by the natural event problems. According to the Wilcoxon sign-rank test results, the ambiguity attitudes from the Ellsberg's urn problem are identical to those from the natural event problems, implying that there are no significant differences between the a-aversion indexes generated by the Ellsberg's urn problem and any of the three experimental groups. One experimental design feature may have directly facilitated a decision-making bias which may have given rise to this result. The Ellsberg's urn problem was always presented first to the respondents and it is reasonable to concede the possibility that this may have led to the emergence of anchoring or insufficient adjustment effect.

Finally, efforts to replicate the competence effect using the primary data gathered for the purposes of this experiment proved unsuccessful. I also attempted (the resulting output is not presented here) to establish the competence effect using the original method proposed in Heath and Tversky (1991). I calculated the median perceived competence score and the mean a-aversion index. I then created a binary variable measuring whether a respondent's perceived competence score is greater than the median and whether they preferred the uncertain over the risky bet (a-aversion index lower than the mean). The resulting binary correlation coefficient is positive but of very modest magnitude and insignificant at the most common significance levels.

6. CONCLUSION

In closing, I believe it is worthwhile to reflect on the key limitations of the present study, some of which may well have been a major reason for the unconvincing empirical results.

First, the experimental design involved measuring the key independent variables using single-item Likert-type questions. As discussed at length in 4.1.1, there is a non-trivial amount of theoretical and quantitative evidence in favour of the use of full-fledged Likert scales for measuring psychological (non-factual) theoretical constructs. The application of an inappropriate measurement technique has serious consequences for the validity of the data. The approach followed here was taken out of practical considerations, so that under a less-constrained scenario it would have been worth it trading off increased complexity for improved robustness of the construct measurement method. A related downside of this is the difficulty of testing the reliability of the answers provided by respondents. Popular approaches such as test-retest or correction for attenuation are highly impractical under the present circumstances and would have made the collection of sufficient number of responses increasingly time-consuming.

It has to be noted as well that the omission of a "don't know "or "no opinion" option presents a further threat to the validity of the data due to the meaningful difference between indifference and ignorance (e.g. see and Grichting, 1994). On this note, one can also make the case that the presentation of the answer options on the competence self-evaluation questions may not reflect the most suitable format given the concept being evaluated. Perhaps respondents were confused by the numerical labels chosen to describe each answer option. For example, the number 0 typically confers neutrality or emptiness and not an average quantity of something (e.g. knowledge), as it would imply here since it is used as the midpoint. Perhaps contributing to the confusion is the fact that respondents were also asked to describe their perceived minimal or non-existent levels of competence by negative numbers and it may be difficult to imagine having a negative level of competence. I anticipated such misinterpretations of the numeric labels. However, I did not choose an answer option spectrum labeled exclusively with positive numbers for the sake of (in addition to the empirical support cited in subsection 3.4) consistency with the desirability measure (which is described much more naturally in terms of positive and negative numeric labels). Inevitably, a few of these limitations are the result of carefully weighed trade-offs.

Second, and a more general critique of Likert-type data, is that it is a significant challenge to keep the “mental benchmarks” fixed across subjects. For example, in the case of competence self-evaluations, it is difficult to argue that if two respondents provide the same score, then it means they indeed should be treated as equally competent. Quite possibly these individuals have different perspectives on what this score means due to their different levels of knowledge (e.g. the more competent you are the more you appreciate how much there is to be learned still) but also due to other motivations. Similar reasoning can be applied to event (un)desirability ratings. Hence, best practice is to give objective benchmarks where possible of what the different scores are equivalent to in practice.

Finally, the potential presence of anchoring of answers on the matching probabilities questions presents another potentially serious limitation of this study. Due to the fixed order in which the Ellsberg’s urn and natural events problems were presented, it may be that respondents were affected by the similarity of the two types of tasks and failed to adjust their answers sufficiently. If such behavior can indeed explain the insignificant differences between the ambiguity indexes generated by the artificial and natural events, then it is possible that this bias may have prevented us from measuring the true differences in attitudes.

To conclude, the limitations of this paper notwithstanding, I believe the ideas and research directions presented here have the potential to add further insights into the study of decisions under uncertainty. Few would stand behind the claim that we have a complete understanding of how we make decisions about uncertain events and, although the quest for establishing a causal link between desire and predictions remains a challenge, desirability bias appears a candidate worthy of being the focus of further research into the components of ambiguity attitudes.

7. BIBLIOGRAPHY

1. Allen, I. E., & Seaman, C. A. (2007). Likert Scales and Data Analyses. Quality Progress, 40(7), 64-65. Retrieved from http://www.asq.org/qualityprogress/index.html

2. Babad, E., & Katz, Y. (1991). Wishful thinking—against all odds. Journal of Applied Social Psychology, 21(23), 1921-1938

3. Balcetis, E. (2008). Where the motivation resides and self-deception hides: How motivated cognition accomplishes self-deception. Social and Personality Psychology Compass, 2/1, 361–381

4. Bar-Hillel, M., & Budescu, D. (1995). The elusive wishful thinking effect. Thinking & Reasoning, 1(1), 71-103

5. Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. Journal of marketing research, 44(2), 175-184

6. Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers?. Journal of Experimental Social Psychology, 66, 20-28

7. Buckley, T., & Sniezek, J. (1992). Passion, preference, and predictability in judgmental forecasting. Psychological Reports, 70(3_suppl), 1022-1022

8. Budescu, D. V., & Bruderman, M. (1995). The relationship between the illusion of control and the desirability bias. Journal of Behavioral Decision Making, 8(2), 109-125

9. Camerer, C., and Weber, M. (1992). Recent Developments in Modeling Preferences: Uncertainty and Ambiguity. Journal of Risk and Uncertainty, V (1992), 325-70

10. Carifio J, Perla R. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. Journal of the Social Sciences 2007;3(3), 106–116.

11. Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. Medical Education, 42, 1150–1152

12. Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. Psychological bulletin, 130(5), 813

13. Churchill, Gilbert A. (1979). A Paradigm for Developing Better Measures of Marketing Constructs. Journal of Marketing Research, 16 (February), 64–73

14. Conover, W. J., & Iman, R. L. (1982). Analysis of covariance using the rank transformation. Biometrics, 715-724

15. Cortina, Jose M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. Journal of Applied Psychology, 78 (February), 98–104

16. Crutzen, R., & Peters, G. J. Y. (2017). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. Health Psychology Review, 11(3), 242-247

17. Curley, S. P., Yates, J. F., & Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. Organizational behavior and human decision processes, 38(2), 230-256

18. Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. Personality & Social Psychology Bulletin, 28, 1379–1387

19. Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. Journal of the Academy of Marketing Science, 40(3), 434–449. doi:10.1007/s11747-011-0300-3

20. Dolbier, C. L., Webster, J. A., McCalister, K. T., Mallon, M. W., & Steinhardt, M. A. (2005). Reliability and validity of a single-item measure of job satisfaction. American Journal of Health Promotion, 19(3), 194-198

21. Drolet, A. L., & Morrison, D. G. (2001). Do we really need multipleitem measures in service research? Journal of Service Research, 3 (3), 196–204

22. Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. Journal of risk and uncertainty, 4(1), 5-28

23. Giardini, F., Coricelli, G., Joffily, M., & Sirigu, A. (2008). Overconfidence in predictions as an effect of desirability bias. In Advances in decision making under risk and uncertainty (pp. 163-180). Springer, Berlin, Heidelberg

24. Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. The Quarterly Journal of Economics, 115(3), 791-810

25. Granberg, D. and Brent, E. (1983). When prophecy bends: The preference–expectation link in US presidential elections, 1952–1980. Journal of Personality and Social Psychology, 45(3), 477

26. Jamieson, S. (2004). Likert scales: how to (ab) use them. Medical education, 38(12), 1217-1218

27. Kilka, M., & Weber, M. (2001). What determines the shape of the probability weighting function under uncertainty?. Management science, 47(12), 1712-1726

28. Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing research, 39(2), 121-123

29. Koehler, J. J., Gibbs, B. J., & Hogarth, R. M. (1994). Shattering the illusion of control: Multi-shot versus single-shot gambles. Journal of Behavioral Decision Making, 7(3), 183-191

30. Krizan, Z., Miller, J. C., & Johar, O. (2010). Wishful thinking in the 2008 US presidential election. Psychological Science, 21(1), 140-146.

31. Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological bulletin*, *133*(1), 95.

32. Krizan, Z., & Windschitl, P. D. (2009). Wishful thinking about the future: Does desire impact optimism? *Social and Personality Psychology Compass*, *3*(3), 227-243.

33. Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), Advances in survey research (pp. 29-44). San Francisco: Sage

34. Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480

35. Lench, H. C. (2009). Automatic optimism: The affective basis of judgments about the likelihood of future events. Journal of Experimental Psychology: General, 138(2), 187.

36. Li, Z., Müller, J., Wakker, P. P., & Wang, T. V. (2017). The Rich Domain of Ambiguity Explored. *Management Science*.

37. Likert, R. (1932). A Technique for the Measurement of Attitudes. Archives of Psychology, Vol. 140, No. 55.

38. Massey, C., Simmons, J. P., & Armor, D. A. (2011). Hope over experience: Desirability and the persistence of optimism. Psychological Science, 22(2), 274-281

39. Marks, R. W. (1951). The effect of probability, desirability, and “privilege” on the stated expectations of children. Journal of Personality, 19(3), 332-351

40. McKenna, F. P. (1993). It won't happen to me: Unrealistic optimism or illusion of control?. British Journal of Psychology, 84(1), 39-50

41. Mckenzie, N., & Marks, I. (1999). Quick rating of depressed mood in patients with anxiety disorders. The British Journal of Psychiatry, 174(3), 266-269

42. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81

43. Morewedge, C. K., Tang, S., & Larrick, R. P. (2016). Betting your favorite to win: Costly reluctance to hedge desired outcomes. Management Science, 64(3), 997-1014

44. Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. Advances in health sciences education, 15(5), 625-632

45. Olsen, R. A. (1997). Desirability bias among professional investment managers: Some evidence from experts. Journal of Behavioral Decision Making, 10(1), 65-72

46. Peter, Paul J. (1979). Reliability: A Review of Psychometric Basics and Recent Marketing Practices. Journal of Marketing Research, 16 (February), 6–17

47. Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach’s alpha and the route towards more comprehensive assessment of scale quality. European Health Psychologist, 16(2), 56-69

48. Price, P. C. (2000). Wishful thinking in the prediction of competitive outcomes. Thinking & Reasoning, 6(2), 161-172

49. Quade, D. (1967). Rank analysis of covariance. Journal of the American Statistical Association, 62(320), 1187-1200

50. Revelle, W. (1979). Hierarchical Clustering and the Internal Structure of Tests. Multivariate Behavioral Research, 14 (1), 57–74

51. Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. Psychological Methods, 17, 354-373. doi: 10.1037/a0029315

52. Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. Journal of personality and social psychology, 95(2), 293

53. Roese, N. J., & Olson, J. M. (2007). Better, stronger, faster: Self-serving judgment, affect regulation, and the optimal vigilance hypothesis. Perspectives on Psychological Science, 2(2), 124-141

54. Rosenthal, R., & Fode, K. L. (1963). Psychology of the scientist V: Three experiments in experimenter bias. Psychological Reports, 12, 491–511

55. Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. International journal of research in marketing, 19(4), 305-335

56. Russo, J. E., & Corbin, J. C. (2016). Not by desire alone: The role of cognitive consistency in the desirability bias. Judgment and Decision Making, 11(5), 449

57. Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there?. Personnel psychology, 36(3), 577-600

58. Stevens, S. S. (1946). On the theory of scales of measurement. Science, New Series, Vol. 103, No. 2684 (Jun. 7, 1946), pp. 677-680

59. Stuart, J. O. R., Windschitl, P. D., Smith, A. R., & Scherer, A. M. (2017). Behaving optimistically: How the (Un) desirability of an outcome can bias people's preparations for it. Journal of Behavioral Decision Making, 30(1), 54-69

60. Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. Journal of Experimental Psychology: General, 146(8), 1143

61. Uhlaner, C. J., & Grofman, B. (1986). The race may be close but my horse is going to win: Wish fulfillment in the 1980 presidential election. Political Behavior, 8(2), 101-129

62. Van Dijk, W. W., Zeelenberg, M., & Van der Pligt, J. (2003). Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment. Journal of Economic Psychology, 24(4), 505-516

63. Vosgerau, J. (2010). How prevalent is wishful thinking? Misattribution of arousal causes optimism and pessimism in subjective probabilities. Journal of Experimental Psychology: General, 139(1), 32

64. Keren, G., & Wagenaar, W. A. (1988). Chance and skill in gambling: A search for distinctive features. Social Behaviour, 3(3), 199-217

65. Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: a replication and extension. Organizational Research Methods, 4(4), 361-375

66. Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: how good are single-item measures?. Journal of applied Psychology, 82(2), 247.

67. Weinstein, N. D. (1980). Unrealistic optimism about future life events. Journal of personality and social psychology, 39(5), 806

68. Windschitl, P. D., Scherer, A. M., Smith, A. R., & Rose, J. P. (2013). Why so confident? The influence of outcome desirability on selective exposure and likelihood judgment. Organizational Behavior and Human Decision Processes, 120(1), 73-86

69. Windschitl, P. D., Smith, A. R., Rose, J. P., & Krizan, Z. (2010). The desirability bias in predictions: Going optimistic without leaving realism. Organizational Behavior and Human Decision Processes, 111(1), 33-47.

8. APPENDIX

Start of Block: Introduction

“*Intro Thank you for taking part in this research project!*
The aim of this questionnaire is to capture your attitudes towards uncertain events, and to investigate possible underlying psychological mechanisms. This area of research is relevant for understanding our decision-making process, as in practice we have to make decisions regarding all kinds of uncertain events.
Your answers will be used solely for the purposes of this research project. You will not be asked for personally identifiable information.
Please send any questions, concerns or general comments to georgi.kondov@student.eur.nl

Thank you once more for contributing to the progress of science!”

End of Block: Introduction

Start of Block: Ellsberg

Ellsberg_1

"*Imagine the following situation. A friend of yours fills a bucket with ninety either **red**, **blue**, or **green** balls. Your friend decides in what proportion these colors will be and keeps this information secret from you.*

Then, you are offered a choice between two bets. Please indicate on each line whether you prefer the bet on the left or the bet on the right":

	1 (1)	2 (2)	
You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)

You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if you pull a **red** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 100% probability

Ellsberg_2

"*Similarly, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if you pull a **blue** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 100% probability

Ellsberg_3

"*Similarly, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if you pull a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if you pull a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if you pull a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if you pull a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if you pull a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if you pull a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if you pull a **green** ball from the bucket (and $0	○	○	You win €100 with 100% probability

otherwise).		

Page Break

Ellsberg_12

"*Similarly, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0 otherwise).	◯	◯	You win €100 with 20% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0 otherwise).	◯	◯	You win €100 with 40% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0 otherwise).	◯	◯	You win €100 with 60% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0 otherwise).	◯	◯	You win €100 with 70% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0 otherwise).	◯	◯	You win €100 with 80% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0 otherwise).	◯	◯	You win €100 with 90% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **blue** ball from the bucket (and $0	◯	◯	You win €100 with 100% probability

otherwise).		

Ellsberg_13

"*Similarly, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 90% probability (and $0 otherwise)
You win €100 if you pull a **red** or a **green** ball from the bucket (and $0	○	○	You win €100 with 100% probability

otherwise).		

Ellsberg_23

"*Similarly, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0 otherwise).	○	○	You win €100 with 90% probability (and $0 otherwise)
You win €100 if you pull a **blue** or a **green** ball from the bucket (and $0	○	○	You win €100 with 100% probability

otherwise).

End of Block: Ellsberg

Start of Block: Control_problem

Control_prob1

"*Now, your friend replaces the three-color bucket bet with the following alternative.*
The new bet concerns how taller or not will the skyscraper in Jeddah (Saudi Arabia) be, compared to the Burj Khalifa in Dubai (UAE). The ***Jeddah Tower*** *is currently under construction and is expected to be completed in 2020.*"

End of Block: Control_problem

Start of Block: Control_matching

Control_1

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the Jeddah Tower ends up **between 0-100 meters taller** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 meters taller** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 meters taller** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 meters taller** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 meters taller** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 meters**	○	○	You win €100 with 80% probability (and $0 otherwise)

taller (and $0 otherwise).			
You win €100 if the Jeddah Tower ends up **between 0-100 meters taller** (and $0 otherwise).	○	○	You win €100 with 100% probability

Control_2

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the Jeddah Tower ends up **between 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200**	○	○	You win €100 with 80% probability (and $0 otherwise)

meters taller (and $0 otherwise).			
You win €100 if the Jeddah Tower ends up **between 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 100% probability

Control_3

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the Jeddah Tower ends up **more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **more than 200 meters**	○	○	You win €100 with 80% probability (and $0 otherwise)

taller (and $0 otherwise).			
You win €100 if the Jeddah Tower ends up **more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 100% probability

Page Break

Control_12

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-200 meters taller** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-**	○	○	You win €100 with 90% probability (and $0 otherwise)

200 meters taller (and $0 otherwise).			
You win €100 if the Jeddah Tower ends up **between 0-100 or 100-200 meters taller** (and $0 otherwise).			You win €100 with 100% probability

Control_13

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters**	○	○	You win €100 with 80% probability (and $0 otherwise)

taller (and $0 otherwise).			
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 90% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 0-100 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 100% probability

Control_23

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters taller** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters**	○	○	You win €100 with 80% probability (and $0 otherwise)

taller (and $0 otherwise).			
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters taller** (and $0 otherwise).	◯	◯	You win €100 with 90% probability (and $0 otherwise)
You win €100 if the Jeddah Tower ends up **between 100-200 or more than 200 meters taller** (and $0 otherwise).	◯	◯	You win €100 with 100% probability

End of Block: Control_matching

Start of Block: Control_comp

Control_comp1

"*How competent or incompetent would you consider yourself in a real-world situation requiring you to judge how taller or not will the Jeddah Tower be compared to the Burj Khalifa?*"

Complete lack of competence						Complete competence
-3	-2	-1	0	1	2	3

Competence scale ()	

Page Break

End of Block: Control_comp

Start of Block: Control_des

Control_des1

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the Jeddah Tower ends up between 0-100 meters taller than the Burj Khalifa?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability scale ()							

Page Break

Control_des2

“*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the Jeddah Tower ends up between 100-200 meters taller than the Burj Khalifa?*”

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability scale ()							

Page Break

Control_des3

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the Jeddah Tower ends up more than 200 meters taller than the Burj Khalifa?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability scale ()							

End of Block: Control_des

Start of Block: Undesirability_problem

Undesirability_prob1

"*Now, your friend replaces the three-color bucket bet with the following alternative.*

The new bet concerns the proportion of today's occupations that will be automatable by 2025. ***Job automation*** *involves the replacement of human workers by technologies such as machine learning and robotics.*"

End of Block: Undesirability_problem

Start of Block: Undesirability_matching

Undes_1

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the proportion ends up **exceeding 30%** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30%** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30%** (and	○	○	You win €100 with 100% probability

$0 otherwise).		

Undes_2

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the proportion ends up **between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30%**	○	○	You win €100 with 100% probability

(and $0 otherwise).		

Undes_3

“*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*”:

	1 (1)	2 (2)	
You win €100 if the proportion ends up **between 0%-15%** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 0%-15%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 0%-15%** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 0%-15%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 0%-15%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 0%-15%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 0%-15%** (and	○	○	You win €100 with 100% probability

$0 otherwise).		

Page Break

Undes_12

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the proportion ends up **exceeding 30% or is between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is**	○	○	You win €100 with 90% probability (and $0 otherwise)

between 15%-30% (and $0 otherwise).			
You win €100 if the proportion ends up **exceeding 30% or is between 15%-30%** (and $0 otherwise).	○	○	You win €100 with 100% probability

Undes_13

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the proportion ends up **exceeding 30% or is between 0%-15%**(and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 0%-15%**(and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 0%-15%**(and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 0%-15%**(and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is between 0%-15%**(and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the proportion ends up **exceeding 30% or is**	○	○	You win €100 with 90% probability (and $0 otherwise)

between 0%-15%(and $0 otherwise).			
You win €100 if the proportion ends up **exceeding 30% or is between 0%-15%**(and $0 otherwise).	○	○	You win €100 with 100% probability

Undes_23

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the proportion ends up **between 15%-30% or 0%-15%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30% or 0%-15%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30% or 0%-15%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30% or 0%-15%** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30% or 0%-15%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the proportion ends up **between 15%-30% or**	○	○	You win €100 with 90% probability (and $0 otherwise)

0%-15% (and $0 otherwise).			
You win €100 if the proportion ends up **between 15%-30% or 0%-15%** (and $0 otherwise).	○	○	You win €100 with 100% probability

End of Block: Undesirability_matching

Start of Block: Undesirability_comp

Undes_comp1

"*How competent or incompetent would you consider yourself in a real-world situation requiring you to judge the proportion of jobs that will be automatable by 2025?*"

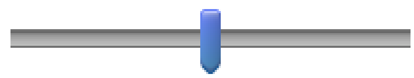
	Complete lack of competence						Complete competence
	-3	-2	-1	0	1	2	3
Competence scale ()							

End of Block: Undesirability_comp

Start of Block: Undesirability_des

Undes_des1

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the proportion of jobs which will be automatable by 2025 ends up exceeding 30%?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability scale ()							

Page Break

Undes_des2

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the proportion of jobs which will be automatable by 2025 ends up between 15%-30%?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability ()							

Page Break

Undes_des3

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the proportion of jobs which will be automatable by 2025 ends up between 0%-15%?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability ()							

End of Block: Undesirability_des

Start of Block: Desirability_problem

Desirability_prob1

"*Now, your friend replaces the three-color bucket bet with the following alternative.*

The new bet concerns the share of global energy production in 2025 that will come from ***renewable sources*** *(solar, wind, geothermal, biomass, hydropower).*"

End of Block: Desirability_problem

Start of Block: Desirability_matching

Des_1

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the share ends up between **0%-15%** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the share ends up between **0%-15%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the share ends up between **0%-15%** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the share ends up between **0%-15%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the share ends up between **0%-15%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the share ends up between **0%-15%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the share ends up between **0%-15%** (and $0	○	○	You win €100 with 100% probability

otherwise).		

Des_2

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the share ends up between **15%-30%** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the share ends up between **15%-30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the share ends up between **15%-30%** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the share ends up between **15%-30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the share ends up between **15%-30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the share ends up between **15%-30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the share ends up between **15%-30%** (and $0	○	○	You win €100 with 100% probability

otherwise).		

Des_3

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the share ends up **greater than 30%** (and $0 otherwise).	○	○	You win €100 with 10% probability (and $0 otherwise)
You win €100 if the share ends up **greater than 30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the share ends up **greater than 30%** (and $0 otherwise).	○	○	You win €100 with 30% probability (and $0 otherwise)
You win €100 if the share ends up **greater than 30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the share ends up **greater than 30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the share ends up **greater than 30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the share ends up **greater than 30%** (and $0	○	○	You win €100 with 100% probability

otherwise).		

Page Break

Des_12

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the share ends up **between 0%-15% or 15%-30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or 15%-30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or 15%-30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or 15%-30%** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or 15%-30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or 15%-30%** (and $0 otherwise).	○	○	You win €100 with 90% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or 15%-30%**	○	○	You win €100 with 100% probability

(and $0 otherwise).		

Des_13

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the share ends up **between 0%-15% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the share ends up **between 0%-15% or greater**	○	○	You win €100 with 90% probability (and $0 otherwise)

than 30% (and $0 otherwise).			
You win €100 if the share ends up **between 0%-15% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 100% probability

Des_23

"*Once more, please indicate on each line whether you prefer the bet on the left or the bet on the right*":

	1 (1)	2 (2)	
You win €100 if the share ends up **between 15%-30% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 20% probability (and $0 otherwise)
You win €100 if the share ends up **between 15%-30% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 40% probability (and $0 otherwise)
You win €100 if the share ends up **between 15%-30% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 60% probability (and $0 otherwise)
You win €100 if the share ends up **between 15%-30% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 70% probability (and $0 otherwise)
You win €100 if the share ends up **between 15%-30% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 80% probability (and $0 otherwise)
You win €100 if the share ends up **between 15%-30% or greater**	○	○	You win €100 with 90% probability (and $0 otherwise)

than 30% (and $0 otherwise). You win €100 if the share ends up **between 15%-30% or greater than 30%** (and $0 otherwise).	○	○	You win €100 with 100% probability

End of Block: Desirability_matching

Start of Block: Desirability_comp

Des_comp1

“*How competent or incompetent would you consider yourself in a real-world situation requiring you to judge the share of global energy production in 2025 that will come from renewables?*”

	Complete lack of competence						Complete competence
	-3	-2	-1	0	1	2	3
Competence scale ()							

End of Block: Desirability_comp

Start of Block: Desirability_des

Des_des1

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the share of global energy production in 2025 which will come from renewables ends up between 0%-15%?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability scale ()							

Page Break

Des_des2

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the share of global energy production in 2025 which will come from renewables ends up between 15%-30%?*"

	Strongly undesirable			Indifference			Strongly desirable
	-3	-2	-1	0	1	2	3
Desirability scale ()							

Page Break

Des_des3

"*In general, and ignoring any bets you may have made so far, how strongly do you wish or do not wish that the share of global energy production in 2025 which will come from renewables ends up greater than 30%?*"

	Strongly undesirable	Indifference	Strongly desirable
	-3 -2	-1 0 1	2 3
Desirability scale ()			

End of Block: Desirability_des

Start of Block: Controls

*

Age: "*Please enter your age*":

____

Gender: "*Please select your gender*":

○ Male (2)

○ Female (3)

Nation: "*Please enter your nationality*":

____

Study: "*Please type the field of study of your latest educational qualification*":

____

End of Block: Controls
