UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

WORKING PAPERS IN ECONOMICS

No 484

Self-Image and Valuation of Moral Goods: Stated versus Real Willingness to Pay

Olof Johansson-Stenman and Henrik Svedsäter

January 2011

ISSN 1403-2473 (print)
ISSN 1403-2465 (online)

Department of Economics
School of Business, Economics and Law at University of Gothenburg
Vasagatan 1, PO Box 640, SE 405 30 Göteborg, Sweden
+46 31 786 0000, +46 31 786 1326 (fax)
www.handels.gu.se info@handels.gu.se

Self-Image and Valuation of Moral Goods: Stated versus Real Willingness to Pay

Olof Johansson-Stenman

Department of Economics, School of Business, Economics and Law,
University of Gothenburg, Sweden
e-mail: Olof.Johansson@economics.gu.se

Henrik Svedsäter

Organisational Behaviour, London Business School, United Kingdom
e-mail: hsvedsater@london.edu

Abstract

Hypothetical bias in stated-preference methods appears sometimes to be very large, and other times non-existent. This is here largely explained by a model where people derive utility from a positive self-image associated with morally commendable behavior. The results of a choice experiment are consistent with the predictions of this model; the hypothetical marginal willingness to pay (*MWTP*) for a moral good (contributions to a WWF project) is significantly higher than the corresponding real *MWTP*, whereas no hypothetical bias is present for an amoral good (a restaurant voucher). Moreover, both the theoretical model and the experimental evidence suggest that also the real *MWTP* for the moral good is biased upwards by being higher within than outside the experimental context.

Key words: Stated-preference methods; choice experiment; hypothetical bias; self-image; non-market valuation; warm glow

JEL-classification: C91, D63, Q5

1. Introduction

What determines people's responses in stated-preference (SP) surveys targeting issues with a perceived ethical dimension, such as valuation of environmental and various other types of public goods? And to what extent can we interpret those responses as being representative of underlying preferences? These questions are crucial from a policy perspective, in particular in the US and an increasing number of European countries, where cost-benefit analysis, often making use of SP methods, is compulsory for all major proposed regulations. Although most researchers probably agree that there is potential scope for overstatement in various kinds of SP studies, no consensus exists whether this is a major problem, nor how hypothetical estimates could or should be calibrated to better represent underlying preferences. Perhaps more importantly, few studies have investigated for which types of goods and under what circumstances hypothetical bias is likely to occur, and why this is the case.

In this paper we develop a theoretical model aimed at explaining variations of hypothetical bias in the literature. Drawing on papers by Andreoni (1989; 1990), Kahneman and Knetsch (1992), Akerlof and Kranton (2000), Brekke et al. (2003), Santos-Pinto and Sobel (2005) and Nyborg and Brekke (2010), the model proposes that people, in addition to the instrumental benefits associated with a good, derive utility from a positive self-image. This, in turn, is influenced by (*i*) the degree to which stated or real behavior coincides with the respondents' ethical views, and (*ii*) the extent to which respondents are honest to themselves. The model predicts that people overstate their marginal willingness to pay *(MWTP)* in SP studies for goods with a perceived ethical dimension, denoted moral goods, but not for morally neutral goods. The model furthermore predicts that also the elicited real-money *MWTP* exaggerates people's valuation of a moral good, although to a lesser extent.

In order to test these predictions, we conduct a choice experiment (CE) assessing people's valuation of what we refer to as a moral and an amoral good, respectively. A CE is

an SP method where the respondents make repeated choices between bundles of goods. The method has been increasingly used to value non-market goods (see, e.g., Louviere et al., 2000; List et al., 2006). The moral good is here represented by a donation to a campaign administered by the World Wildlife Fund (WWF) to help save the Asian Elephant, and the amoral good is a voucher valid at a local Italian restaurant in Gothenburg, Sweden. The CE is then compared with the outcome of a similar exercise, based on another but similar sample drawn from the same underlying student population, only this time using real instead of hypothetical monetary trade-offs. The empirical results are consistent with the predictions of our model; the stated *MWTP* for the moral good (the WWF campaign) is significantly higher than the corresponding real *MWTP*, whereas no difference is found between stated and real WTP for the amoral good (the restaurant voucher). In following up on these findings, we illustrate how also the real-money CE exaggerates people's valuation of the moral good, in the sense that the experimental situation *per se* seems to induce a positive bias.

Section 2 presents a brief review of hypothetical bias in SP studies and of relevant psychological and behavioral economics literature that helps to explain past empirical results. Section 3 presents a formalized model and derives testable hypotheses, whereas Section 4 outlines the CE design for assessing the value of our moral and amoral good. The empirical results are presented in Section 5, while Section 6 examines to what extent the real-money CE is representative of the value people in reality place on actual changes of the moral good outside the experimental context. Section 7 discusses the findings in a broader context.

2. Literature Review

2.1 The Existence of Hypothetical Bias

The extent to which *WTP* statements correspond with actual payments is often seen as the ultimate validity test of SP methods. List and Gallet (2001) and Murphy et al. (2005)

conducted meta-studies on observed disparities between hypothetical and actual *WTP* in contingent valuation (CV) studies, reporting that hypothetical *WTP* generally exceeds actual *WTP*, and that the difference tends to be larger for public than for private goods. Murphy et al. (2005) also found a much lower hypothetical bias in studies that relied on a within-subject test of hypothetical and actual *WTP* than in studies making split-sample comparisons between subjects.[1]

However, some other studies report no statistically significant differences between hypothetical and actual *WTP*. Of particular relevance to our work are Carlsson and Martinsson (2001) and Cameron et al. (2002), who used CEs to value what we here denote moral goods. In Carlsson and Martinsson (2001), the respondents first made 16 hypothetical pair-wise choices and then 16 similar (but not identical) pair-wise choices with real-money implications. No significant difference was found between hypothetical and actual marginal *WTP* for donations to a variety of environmental projects, although the former was 10-15 percent higher than the latter. Cameron et al. (2002) tested several elicitation formats in a comprehensive study and found that the mean *WTP* was between 30 and 330 percent larger in hypothetical CEs. However, due to large error terms, a common underlying preference structure could not be rejected.[2]

For some environmental goods, such as access to recreation sites or hunting rights, it is possible to compare SP methods with revealed preference (RP) methods, for instance by using travel-cost or hedonic-pricing methods. In a meta-analysis by Carson (1996), values obtained from RP studies were found to be of the same order of magnitude as those derived using dichotomous-choice CV studies. Risk and time valuations are other examples where

[1] For a direct test of hypothetical bias in within- and between-subject designs, see Johansson-Stenman and Svedsäter (2008).

both SP and RP methods are routinely used. In another large meta-analysis, Kochi et al. (2003) found that CV studies on average result in significantly lower values of statistical lives than studies based on the hedonic-pricing method. Finally, Wardman (2001) performed a meta-study of British value-of-time studies and found relatively small differences, although SP studies on average yielded somewhat lower values.

2.2 Possible Explanations behind Hypothetical Biases

The most frequently assumed reason for a positive hypothetical bias is that respondents simply do not take hypothetical questions seriously. However, if this assumption were correct, we would expect to see a greater variance of bids and not a systematic bias upwards. Given the above empirical pattern and the high policy relevance, it appears worthwhile to investigate more systematically for which goods and under what circumstances hypothetical statements are likely to be biased, and when they are not. Moreover, we would like to have an intuitively plausible theory for *why* overstatements frequently occur in some contexts but not in others.

There is much evidence from psychology, and more recently from behavioral economics, that people like to have a positive self-image, and that they try to maintain this image in various ways (Gilovich, 1991; Baumeister 1998). Consistent with this, it has been found that most people believe that they perform a variety of tasks better than the average person (e.g., perceiving themselves as better drivers, or that they are smarter). Central to our argument here is that moral identity is part of an individual's self-image (e.g., Aquino and Reed II, 2002), which is often associated with certain beliefs, attitudes, and behaviors (Shih et al., 1999; Forehand et al., 2002). The fact that people prefer to see themselves as more socially responsible than others (e.g., Gilovich, 1991; Taylor and Brown, 1994) only serves to

[2] Presumably, this was partly due to the fact that there was much less variation in the real-money bids; cf. Carlsson and Johansson-Stenman (2010).

illustrate the importance people attach to moral identity. In this context, Johansson-Stenman and Martinsson (2006) asked people about what characteristics they considered to be important when buying a car. Whereas most claimed environmental characteristics to be very important, very few emphasized the status associated with a specific car or model. Interestingly, when asked about what characteristics they believed were important *for others*, the reverse pattern emerged insofar as status became much more important and environmental aspects less important. This underlines the desirability of this trait, and the tendency to view oneself 'better' than others in this respect. Similar findings are reported by Brekke et al. (2003). When investigating people's motivation behind recycling in a Norwegian survey, as many as 73% of the respondents answered that one of their main reasons was that they would like to see themselves as responsible citizens.

Provided that a high *WTP* is seen as something honorable, and hence enhancing a person's self-image, it follows that people have an incentive based on self-deception to overstate their *WTP*. This obviously applies to both stated and actual *WTP*, but since nothing needs to be paid in a hypothetical context, a positive hypothetical bias is logical. Our reasoning also corresponds with psychological theories arguing that people derive value from merely expressing certain opinions or attitudes (e.g., Katz, 1960; Herek, 1986), particularly under circumstances when these are not binding or directly tied to outcomes (Kahneman and Knetsch, 1992; Bodner, 1995). Hence, when a verbal statement is free of charge, more emphasis will be placed on maintaining a positive self-image than when economic costs are involved. The self-image motive proposed here also helps explain the observed pattern of higher hypothetical bias for public goods, since these goods often have moral implications. Arguably, the preference for saving wild animals from extinction is built on different

premises and is more strongly associated with a moral code of conduct than, say, private access to fishing or hunting rights or consumption of chocolate bars.[3]

However, since we do not observe infinite *WTP*s in hypothetical SP studies even for clearly ethical issues, some moderating factor must play a role. Our presumption is that people simultaneously want to be honest to themselves, knowing that there is a limit as to how much they are able or willing to commit. Assuming that you won 100,000 USD in a lottery, how much of this would you donate to charity? Even if we agree that the most honorable thing to do would be to donate all of it, most of us would not, and we would probably feel dishonest if claiming to do that in a hypothetical survey. This motive can also contribute to our understanding of why so-called cheap-talk scripts tend to result in lower hypothetical bias, since they aim to make respondents more honest and realistic in their answers (see, e.g., Cummings and Taylor, 1999, and List et al., 2006, for CV and CE applications of such scripts).

It is worth emphasizing that the self-image motive proposed here should not be confused with preference falsification or with the willingness to impress or provide informative signals to other people (Bernheim, 1994; Kuran, 1995; Neilson, 2009). List et al. (2004), for example, showed that CV respondents are much more willing to vote in favor of a costly environmental project if others are informed about their choice. Here the assumed driving force can instead be seen as *self*-signaling (Bodner and Prelec, 2003), implying that opinions and actions provide signals to ourselves as to what kind of person we are, including our intentions toward the matter at stake. Like Adam Smith (1759) and Benabou and Tirole

[3] Hypothetical bias for purely private goods such as chocolate bars or sunglasses can obviously not be explained by the self-image effects discussed here (e.g., Cummings et al., 1995; Lusk and Schroeder, 2003). One plausible interpretation of such findings is that some respondents are actually answering a slightly different question to the one being asked. In order to make sense of the inquiry raised they may, for example, ask themselves "How much would I be willing to pay if I were to buy a pair of sunglasses today?"

(2006), we may think of an individual who makes moral decisions by assessing his/her own conduct from the perspective of how an ideal person would act in a certain situation, independent of whether or not his/her actions are being observed. The assumed mechanism is therefore effective also in highly anonymous contexts, as in the experiment conducted here, and does not rely on whether actions are publicly known.[4]

3. The Theoretical Model

Our model can be seen as an extension of Andreoni's (1989, 1990) model, where people derive a "warm-glow" from contributing to a "good cause" (which public goods are often seen as), and of the idea developed by Kahneman and Knetsch (1992) that people's value statements in SP surveys represent the "purchase of moral satisfaction." A key feature of these models is that people gain utility intrinsically from their own contributions but not from those of others. The extensions suggested here consist of specifying why and when people receive such a warm-glow, and of considering the limits posed by people's contention to be honest. Consider a strictly increasing and quasi-concave utility function as follows:

$$U = u(Money, Rest, WWF), \tag{1}$$

where *Money* is private consumption (or income), and *Rest* and *WWF* represent money for a restaurant voucher and a WWF campaign, respectively. The true marginal willingness to pay for *Rest* in terms of *Money* is then given by:

$$MWTP_{Rest}^{true} \equiv -\left.\frac{\mathrm{d}Money}{\mathrm{d}Rest}\right|_{u} = \frac{\partial u}{\partial Rest} \Big/ \frac{\partial u}{\partial Money}, \tag{2}$$

[4] For recent applications and theorizing of similar motivational sources in economics, see for example Murninghan et al. (2001) and Benabou and Tirole (2002, 2004, 2006). Likewise; for measures of the relative strength of extrinsic versus intrinsic motives of voluntary acts, see Alpizar et al. (2008) and Lacetera and Macis (2010).

i.e., the marginal rate of substitution between *Rest* and *Money*. Similarly, the true marginal willingness to pay for *WWF* in terms of *Money* is given by:

$$MWTP_{WWF}^{true} \equiv -\frac{\mathrm{d}Money}{\mathrm{d}WWF}\bigg|_{u} = \frac{\partial u}{\partial WWF} \Big/ \frac{\partial u}{\partial Money}. \tag{3}$$

In most SP experiments aimed to measure the value of a change in a public good, respondents are assumed to maximize a function *u* as above, so that $MWTP^{stated} = MWTP^{true}$.

By including self-image effects, *s*, into the model (following Akerlof and Kranton, 2000, 2002; Brekke et al., 2003; Santos-Pinto and Sobel, 2005; Johansson-Stenman and Martinsson, 2006; Alpizar et al., 2008; and Brekke and Nyborg, 2010), we have instead the following utility function:

$$V = v\big(u(Money, Rest, WWF), s\big), \tag{4}$$

where $\frac{\partial v}{\partial s} > 0$. It is here assumed that people's utility, in addition to changes of *Money* and *WWF*, depends on how their self-image is affected by their intentions and actions. An individual's self-image is of course influenced by many factors. In the model proposed here, *s* is assumed to depend on (*i*) the degree to which individuals act in accordance with their ethical beliefs (*ethics* for short), and (*ii*) the extent to which they are honest to themselves (*honesty* for short). Hence:

$$s = f(d_{Rest}^{ethics}, d_{Rest}^{honesty}, d_{WWF}^{ethics}, d_{WWF}^{honesty}), \tag{5}$$

where $d_i^{ethics} \equiv \left| MWTP_i^{stated} - MWTP_i^{moral} \right|$ is the absolute value of the difference between the stated *MWTP* for good *i* and its corresponding ethically superior value, $MWTP_i^{moral}$ (i.e., the value that would maximize the respondent's self-image should there be no conflicts with other determinants of self-image); $d_i^{honesty} \equiv \left| MWTP_i^{stated} - MWTP_i^{true} \right|$ is similarly the absolute value of the difference between stated and true *MWTP*, where the latter is defined as the resulting

MWTP when holding *s* constant (or, equivalently, the amount of money that could be taken away from the individual per dollar given by *someone else* to the good). Let us also assume that $\frac{\partial f}{\partial d_i^{ethics}} < 0$ for $d_i^{ethics} > 0$, $\frac{\partial f}{\partial d_i^{ethics}} = 0$ for $d_i^{ethics} = 0$, $\frac{\partial f}{\partial d_i^{honesty}} < 0$ for $d_i^{honesty} > 0$, $\frac{\partial f}{\partial d_i^{honesty}} = 0$ for $d_i^{honesty} = 0$, $\frac{\partial^2 f}{\partial (d_i^{ethics})^2} < 0$, $\frac{\partial^2 f}{\partial (d_i^{honesty})^2} < 0$, where the shape of the second derivatives ensures a unique optimum. Thus, in a survey of a particular good, a statement deviating from both the morally superior value and the true value causes a disutility for the individual. In our case we have two goods to be valued, so that

$$s = f\left(\left|MWTP_{Rest}^{stated} - MWTP_{Rest}^{moral}\right|, \left|MWTP_{Rest}^{stated} - MWTP_{Rest}^{true}\right|, \left|MWTP_{WWF}^{stated} - MWTP_{WWF}^{moral}\right|, \left|MWTP_{WWF}^{stated} - MWTP_{WWF}^{true}\right|\right) . \quad (6)$$

Let us further define a good *i* to be a moral good if and only if $MWTP_i^{moral} > MWTP_i^{true}$. Correspondingly, a good *i* is defined to be amoral if and only if $MWTP_i^{moral} = MWTP_i^{true}$.[5] Assuming money to the WWF campaign to be a moral good and the restaurant voucher to be an amoral good, we are able to derive the following hypotheses (see Appendix for proofs):

HYPOTHESIS 1. *Given that money to the WWF campaign constitutes a moral good, we have:* $MWTP_{WWF}^{true} < MWTP_{WWF}^{real} < MWTP_{WWF}^{hyp} < MWTP_{WWF}^{moral}$.

HYPOTHESIS 2. *Given that the restaurant voucher constitutes an amoral good, we have:* $MWTP_{Rest}^{true} = MWTP_{Rest}^{real} = MWTP_{Rest}^{hyp} = MWTP_{Rest}^{moral}$.

[5] It is then natural to define a good as an *immoral* good if and only if $MWTP_i^{moral} < MWTP_i^{true}$, although we will not make use of this case in the current paper.

Thus, the model predicts that hypothetical *MWTP* exceeds real-money *MWTP* for goods with moral implications, but not otherwise. An interesting feature of the model is that it also calls into question the validity of real monetary trade-offs assessed in an experimental context and the potential problems that arise for cost-benefit analysis, since the model predicts that $MWTP_{WWF}^{true} < MWTP_{WWF}^{real}$. The intuition is as follows: Consider first an individual who obtains a positive self-image from making a monetary contribution toward an environmental improvement in an experimental context. Consider next the same individual, but this time facing a compulsory tax increase that is combined with an identical environmental improvement. Would he/she then also obtain the same self-image benefit? Presumably not since this is a forced outcome, implying that experimentally elicited values would lead to an overestimation of the benefits most relevant for cost-benefit analysis. We will return to this issue in Section 7.

4. Experimental Design

The group of subjects consisted students at Gothenburg University, enrolled in a wide range of different undergraduate and graduate courses. They were recruited from a pool of students volunteering at the beginning of each semester to participate in experiments run by the university. Altogether 160 students chose to participate in the experiment reported here, which was conducted in individual sessions with one subject at a time. The average age of the subjects was 26.84 years (SD=7.59 years), and 56 (35.0%) were men and 104 (65%) women.

The design is largely based on that used by Carlsson and Martinsson (2001), with two important exceptions: (ii) we use two different goods, one moral and one amoral, and (ii) we use a between-subject, rather than a within-subject design in order to prevent the respondents from being influenced by their previous hypothetical responses in the real money treatment (cf. Johansson-Stenman and Svedsäter, 2008). The subjects were randomly divided into two

sub-samples. Subjects in one sub-sample made hypothetical choices, whereas subjects in the other made choices involving real monetary trade-offs.

The sessions started with two questions about gender and age. The subjects then received verbal and written instructions about the choice experiment, and were informed that its main purpose was to assess the value people place on various goods and services, in this case a voucher valid at a local restaurant and a donation to the WWF to help save the Asian Elephant.[6] Accordingly, they would later be presented a number of trade-offs associated with these goods. The characteristics of the restaurant and the features and purposes of the WWF campaign were explained. In the hypothetical setting, the subjects were instructed that they should answer the questions as truthfully as possible, carefully taking into account how much each good was actually worth to them and how much they could afford to contribute. They were also informed that hypothetical valuation of environmental goods and services is commonly used as a means to inform public policy making, thereby emphasizing the importance of providing realistic answers. The instructions to the subjects are presented in Appendix B.

There were five cash payment levels offered to the subjects (SEK 0, 40, 80, 120, 160), and five levels of the donation or voucher value (SEK 0, 60, 120, 180, 240). The specific amounts were chosen on the basis of a pre-test carried out prior to the main study, involving 20 respondents. Altogether 32 unique choice sets were constructed from these amounts, 16 that valued a restaurant visit and 16 a WWF donation. The 32 choice sets were divided into two blocks of choice sets, Block A and Block B. Each subject was faced with 16 choice sets (i.e., either with Block A or Block B). Eight choice sets in each block concerned a trade-off

[6] The specific restaurant used in this study is an Italian mid-priced restaurant that the vast majority of the students at the university are familiar with.

between cash payment and a restaurant voucher, and the remaining eight a trade-off between a cash payment and a donation to the WWF campaign.

Finally, the order of the choice sets within each block was varied by either presenting eight choice sets valuing the restaurant visit first and then eight valuing the WWF campaign, or vice versa. The two blocks and the different orders of choice sets were balanced across the two sub-samples, hence occurring with the same frequency in the hypothetical and real treatment. In order to minimize any potential differences in behavior between men and women, subjects were divided across samples so as to achieve identical gender ratios in the hypothetical and real scenarios (28 men and 52 women in each sample).

In each choice-set, the subjects were asked to choose between two alternatives, A and B. In eight choice sets, each alternative specified the amounts of money that the subject and the WWF would receive. In the other eight choice sets, each alternative instead specified how much money the subject would receive, and the value of the restaurant voucher. The subjects were specifically told to make each choice independently of the others, and were informed that there were no direct associations between them. Figure 1 below presents two examples of the choice sets used. The complete list of choice sets can be found in Appendix C.

INSERT FIGURE 1 ABOUT HERE

Subjects in the hypothetical treatment were compensated with a fixed show-up fee of SEK 50. The participants in the real treatment, on the other hand, were prior to the task informed that the actual monetary payoffs and donation or restaurant voucher would be based on one of the sixteen choice sets; after completing the task, a number from 1 to 16 would be randomly drawn under the supervision of the participant. This number would specify which choice set would decide the actual payoffs. For example, if number 2 were drawn, then the choice made by the subject in the second choice set counting from the beginning of the questionnaire would determine his/her cash payment, and similarly the value of the restaurant voucher or

the size of the WWF donation. In this way it was in the subjects' interest to treat each choice in isolation, i.e., as if it were the only choice to make.

Before placing the questionnaire in an envelope, the subjects had the opportunity to check how much money they were to receive and whether a restaurant voucher was to be issued or a donation was to be made (and the exact amount of this). They were also informed that the actual donation would be administered and sent to the WWF by the research team after the data collection was completed. A double-blind procedure was finally used in order to ensure respondent anonymity. Each subject was given a ticket with the same number as printed on the envelope containing his/her questionnaire. The sealed envelopes were then transported to a department secretary who did not know anything about the purpose of the experiment or about who each subject was. This person opened the envelope, checked how much money was owed, and put cash and potentially a restaurant voucher in another envelope with the same number written on it. The subjects were then given a date when they were due to exchange their tickets for the envelopes containing their compensation.

Note that there are no direct free-riding problems associated with our design with respect to measuring preferences for public goods. In contrast, it is widely claimed in the environmental economics literature that people tend to overstate their true WTP because they want to enjoy the benefits of the public good, whether or not having any intention of actually paying for this, if they believe that they would not have to pay (their share) for the good in reality. Yet, in this study the subjects know that if money to the WWF-projects is to be increased, they will have to pay for it.

5. Descriptive Results

Figure 2 and 3 below display the relative frequency of choices favoring a restaurant voucher and a donation to the WWF campaign over cash payment, summarized across all subjects.

The horizontal axes indicate the value of the restaurant voucher and the size of the donation to the WWF (in SEK), respectively. The frequencies represent the average over all levels of trade-offs in cash payments.

As expected, the greater the value of the restaurant voucher or the WWF donation for all levels of cash payments taken together, the more often this alternative is chosen. More importantly for our purposes, apart from the SEK 180 level where the restaurant voucher is slightly (but not significantly) more favored in the hypothetical than in the real scenario, the frequency of choosing the amoral good is roughly the same in the hypothetical and in the real setting across all levels of voucher value. For the moral good, on the other hand, the frequency is always greater in the hypothetical than in the real-money context.

<<Insert Figures 2 and 3 about here>>

Chi-square tests indicate that the relative frequency of hypothetical choices favoring the WWF campaign over cash payment is significantly higher than equivalent choices made in the real-money context for all but one donation level ($p<0.01$ for WWF donations of SEK 120 and 240, $p<0.05$ for a donation of SEK 180, and $p=0.11$ for a donation of SEK 60), thus broadly confirming our hypotheses.

6. Econometric Analysis

In the econometric analysis we rely on a standard random-utility framework, assuming that each individual has a utility function consisting of a systematic part, *V*, and a random unobservable term, ε. The utility derived for individual *i* from choosing a given alternative, say Alternative 1, therefore becomes:

$$u_{i1} = V_{i1} + \varepsilon_{i1} \quad . \tag{7}$$

The probability that *i* chooses Alternative 1 then equals the probability that the utility from this alternative is greater than the utility of Alternative 2; i.e.,

$$\Pr(A_i = 1) = \Pr(V_{i1} + \varepsilon_{i1} > V_{i2} + \varepsilon_{i2}) = \Pr(\varepsilon_{i1} - \varepsilon_{i2} > V_{i2} - V_{i1}), \quad (8)$$

where the differences between the error terms in (8) are assumed to be logistically distributed. The systematic part of the utility function, associated with either alternative, is assumed to be linear in the attributes in the interval considered:

$$V_i = \alpha + \beta Money_i + \rho\, Rest_i + \gamma\, WWF_i + \lambda\, HYP\, Rest_i + \mu\, HYP\, WWF_i, \quad (9)$$

where $Money_i$, $Rest_i$, and WWF_i represent money to the respondent, money to the respondent in the form of a restaurant voucher, and money to the WWF project resulting from choosing alternative 1. HYP is a dummy variable that takes the value 1 in the hypothetical treatment and 0 in the real-money treatment. Given the assumed error distribution, the parameters associated with this model (except for the intercept which cancels out) can be estimated with a logit model (see, e.g., Louviere et al., 2000, for a good state-of-the-art overview of the analysis of stated choice models).

Since each individual makes 16 choices in total, the statistical observations are not independent, implying that the standard errors of a basic logit regression, as outlined above, are biased downwards. We deal with this in two different ways; by clustering the error terms at the individual level (using the cluster-command in Stata), and by utilizing Random Effects Model and Fixed Effects Models.[7]

Table 1 below presents the parameter estimates of pooled regression models. The results are in the expected direction according to our theoretical predictions. The parameters associated with money given to subjects, a restaurant voucher, and a donation to WWF,

[7] Where a choice favoring the good over money is always coded "1" and a choice favoring money over the good is always coded "0."

respectively, are all positive and significant at the 0.01 level across all models. Moreover, the interaction effect between hypothetical treatment and a WWF donation is significant at the 0.01 level. This implies that, across all models, the likelihood of trading off cash payment in favor of a donation to the WWF is significantly greater in the hypothetical than in the real-money context. Conversely, no significant difference between the hypothetical and real treatment is found for the restaurant voucher.

<<Insert Table 1 about here>>

However, we are not primarily interested in the logit parameters per se but in the corresponding *MWTP* for each good across treatments. An individual's *MWTP* for an additional dollar given to the WWF campaign in the real-money treatment is given by

$$MWTP_{WWF}^{Actual} = \frac{\partial u_i / \partial WWF}{\partial u_i / \partial Money} = \frac{\partial V_i / \partial WWF}{\partial V_i / \partial Money} = \frac{\gamma}{\beta}. \tag{10}$$

The corresponding *MWTP* in the hypothetical treatment is given by $MWTP_{WWF}^{Hyp} = (\gamma + \mu)/\beta$. Therefore, our measure of hypothetical bias, i.e., the *MWTP* difference between the two treatments, is given by

$$MWTP_{WWF}^{Hyp} - MWTP_{WWF}^{Actual} = \frac{\mu}{\beta}. \tag{11}$$

Similarly, the real-money *MWTP*s for an additional dollar in the form of a restaurant voucher is given by

$$MWTP_{Rest}^{Actual} = \frac{\rho}{\beta}, \tag{12}$$

whereas the corresponding *MWTP* in the hypothetical treatments is $MWTP_{Rest}^{Hyp} = (\rho + \lambda)/\beta$. Hence, the *MWTP* difference for the restaurant voucher between the two treatments is given by

$$MWTP_{Rest}^{Hyp} - MWTP_{Rest}^{Actual} = \frac{\lambda}{\beta}. \tag{13}$$

Table 2 presents the *MWTP*s corresponding to equations (11) to (13), where the standard errors are calculated using the delta method.

<<Insert Table 2 about here>>

The (sample mean) real *MWTP* for an additional dollar in the form of a restaurant voucher is hence equal to 0.54 dollars, whereas it is 0.47 dollars for an additional dollar donated to the WWF campaign. Yet, our main concern here is the difference between the two treatments. For the restaurant voucher, we only have a small (about 0.06) and non-significant hypothetical bias.[8] For the donation to the WWF campaign, in contrast, there is a sizable and statistically significant hypothetical bias of about 0.19 ($p < 0.01$ in all models). This means that the hypothetical *MWTP* for a donation to the WWF is approximately 0.66, or about 40 percent larger than the corresponding real money *MWTP*.

The results are thus consistent with our theoretical model: There is a substantial and statistically significant hypothetical bias for the moral good, and no significant hypothetical bias for the amoral good.

7. Can We Trust the Real-Money Choice Experiment?

The previous section showed that the hypothetical *MWTP* was substantially larger than the corresponding real *MWTP* for a moral good. In this section we will consider another part of HYPOTHESIS 1, derived in Section 3, namely that the *MWTP* estimates from our real-money choice experiment are biased upwards in the sense that an *MWTP* obtained in a real-money experiment tends to exceed the same *MWTP* assessed outside the experimental context, i.e. that $MWTP_{WWF}^{true} < MWTP_{WWF}^{real}$.

[8] This is also true based on a logit model without clustering (not shown), i.e., without taking into consideration the fact that we use a repeated measures design.

We do not attempt to test this part of HYPOTHESIS 1 directly, i.e. by measuring the underlying true MWTP outside the experimental context ($MWTP_{WWF}^{true}$) and comparing it with our measured real money MWTP ($MWTP_{WWF}^{real}$). Instead, by placing the results from the real-money CE in the context of conventional economic theory, it is possible to assess whether $MWTP_{WWF}^{real}$ is a reasonably measure of $MWTP_{WWF}^{true}$ or not.

7.1 A Thought Experiment that Evaluates the Measured Real Money MWTP[9]

In our case, the mean *MWTP* in the real-money treatment is 0.47, implying that subjects at the margin are indifferent between receiving 0.47 dollars themselves and that 1 dollar is given to the WWF project. Now consider Joe Average, who has similar characteristics as the average subject of the experiment conducted here. One day he reads in the newspaper that the save-the-elephants WWF project has experienced a sudden budget cut of 100,000 USD. The next day he is informed that he has won 45,000 USD on a lottery ticket that he bought some weeks ago. As he receives this good news he is presented with an option: A well-known billionaire (for whom Joe doesn't care for) has promised to put in 55,000 USD if Joe puts in 45,000 USD to the WWF save-the-elephants project; the donation would be anonymous and no one ele would know about it. Would he accept this offer? Our conjecture is that he probably would not.

Yet, assuming that Joe acts to maximize the utility function presented in Equation (1), he would actually obtain a higher utility by accepting it. It is important to note that for this we do not need to assume linearity or anything else regarding the structure of the utility function

[9] See Section 9 in Harrison and List (2004) for a discussion of the relations between thought experiments and other kinds of experiments in economics.

beyond (weak) quasi-concavity between *Money* and *WWF*. To see this, note that we formally have that

$$u(Money^0 + L, WWF^0 - 100,000) = u(Money^0, WWF^0), \tag{12}$$

where $Money^0$ is Joe's initial income, and WWF^0 is the WWF elephant budget before the budget cut. L is the minimum amount of money won in the lottery by Joe (or the compensating variation) that would keep him at the same utility level as before the WWF budget cut. Alternatively, we can express L as

$$L = -\int_{WWF^0-10,000}^{WWF^0} \frac{\mathrm{d}Money}{\mathrm{d}WWF}\bigg|_{u=u^0} \mathrm{d}WWF \geq -100,000 \; \frac{\mathrm{d}x(Money^0, WWF^0)}{\mathrm{d}WWF}\bigg|_{u=u^0} = 47,000, \tag{13}$$

where we have taken into consideration the fact that the slope of the indifference curve is most flat in the interval (or at least never flatter than) at $(Money^0, WWF^0)$, which in turn follows directly from the quasi-concavity assumption (i.e. convex indifference curve towards the origin). Since 47,000 is larger than 45,000, Joe should accept the offer. This is perhaps even more straightforward to see graphically. Consider Figure 4 below.

INSERT FIGURE 4 ABOUT HERE

At $(Money^0, WWF^0)$, we have that the slope of the indifference curve reflects the *MWTP*. A budget cut of 100,000 USD for the WWF project then implies that Joe ends up at point *B*. In order to return to the initial utility level, represented by the drawn indifference curve, Joe would have to be indifferent between 100,000 USD to the WWF project (implying a return to *A*) and *L* to himself (implying that he would end up at *D*). Clearly, if he receives 47,000 USD he ends up at *C*. By the quasi-concavity assumption we have that the indifference curve is convex toward the origin, implying that *EV* is larger than 47,000 USD, which in turn is larger than the 45,000 USD won in the lottery.

There is another, perhaps even more important, reason why one should be careful when generalizing the real-money estimates outside the experimental context; the discrepancy

between experienced utility and decision utility (Kahneman et al., 1997; Kahneman and Thaler, 2006). In cost-benefit analysis we are intrinsically concerned with the implications of certain measures in terms of their welfare effects, i.e., the effects in terms of experienced utility. This means that warm-glow effects (Andreoni 1989, 1990) that arise solely in the experimental situation should not be included in cost-benefit analysis. This is because people in general would not experience any warm-glow feelings should the considered measures be implemented by the government, or otherwise not being tied to any personal contributions.

7.2 A Modified Thought Experiment Without Warm-Glow Effects

In order to illustrate the welfare effects with the warm-glow effects removed, consider the following modified thought experiment: As in the previous example, Joe reads in the newspaper that the save-the-elephants WWF project has experienced a 100,000 USD budget cut. The next day one of two mutually exclusive events takes place: A) He reads in the paper that he has won 45,000 USD on the lottery ticket bought earlier. B) He reads that an anonymous billionaire has donated 100,000 USD to the WWF project. Now the question is: Would Joe be better off, in terms of his well-being, in A or in B? Here we have an even stronger reason to believe that A is the correct answer, depite that this, again, is inconsistent with a utility function as in Equation (1).

Thus, we argue that people's *MWTP* in the real-money context of a CE need not be a good indicator of the true value people place on an improvement *per se*. Not only do hypothetical experiments fail to correctly estimate individual welfare effects when important ethical values are involved, the same appears to be true also for real-money experiments, although to a lesser extent. Note that we are not arguing that warm-glow effects from contributing to a good cause *per se* should be excluded from social welfare analyses. On the contrary, we believe that warm-glow feelings are as valid as other motives. However, it is not

appropriate to generalize findings that arise solely in experimental or survey situations[10] to other valuation contexts.[11] That is, if the moral satisfaction occurs primarily when responding to survey questions, or when acting in an economic experiment, then those who are not included in the sample, who obviously constitute the vast majority of the population, would not enjoy this welfare improvement; see also Andreoni (2006, Section 4) for an insightful discussion of whether and when warm-glow effects should be included in social welfare analysis. One possible reason for the apparently very large degree of observed non-selfish behavior is that experimental situations may induce people to think in terms of "what kind of person am I?" to a greater extent than they may otherwise do; cf. Levitt and List (2007) and Akerlof and Kranton (2000).

8. Discussion and Conclusions

It is often argued in the environmental valuation literature that people will reveal their true preferences unless they have a strategic incentive not to do so. However, as argued by Cummings et al. (1997), such "epsilon truthfulness" is a very strong assumption for which there is not much empirical support. Indeed, a frequent criticism of SP methods is that, due to their hypothetical nature, such approaches are likely to result in overestimation of the true values people place on public goods. However, the empirical results of such tests differ, and from several reviewed meta-studies it appears far from correct to conclude that hypothetical survey methods always end up overstating the benefits of public goods.

[10] See Nunes and Schokkaert (2003) for a method to remove warm-glow from a CV survey in order to obtain a "cold" WTP measure.

[11] Still, it is of course possible to argue that there may be benefits that should be taken into account beyond human well-being, and that animal welfare, and perhaps also the environment, should be valued intrinsically (e.g., Singer, 1974).

As far as we know, the model developed here is the first aimed at explaining the observed variation of hypothetical bias across studies. It is also, to our knowledge, the first attempt to test differences between hypothetical and real WTP for public and private goods within the same experimental context. The central tenet of our model is that people derive utility from a positive self-image, which depends on the degree to which they act in accordance with their ethical beliefs and how honest they are to themselves. Thus, people have an incentive, through self-deception and self-signaling, to overstate their true *MWTP*s if a high value is in accordance with their ethical views, but not otherwise. The empirical results presented in this paper are consistent with these hypotheses, and inconsistent with the conventional model typically assumed in the environmental valuation literature; the hypothetical *MWTP* is significantly higher than the real *MWTP* for a public good with moral implications, whereas no such difference is found for a private, morally neutral good.

The model also corresponds with important psychological insights as well as recent findings in the behavioral and experimental economics literature. It draws upon well-established arguments that attitudinal statements and actions do not only reflect instrumental motives, but rest also on presentational concerns and underlying ideals toward which a person aspires. The fact that attitudes and actions partly represent symbolic expressions raises some questions about what should be accounted for in benefit assessment and what should not. Some authors argue that it is irrelevant whether people's preferences reflect selfish interests, instrumental considerations, moral judgements, or any other reasons for that matter, hence suggesting that all such value foundations are valid. Yet, a major problem arises for public policy analysis if these motivations are context specific and do not transcend from one situation to another.

Another feature of our model is that it suggests that even real-money CEs tend to exaggerate people's valuations of moral goods. This is in line with doubts expressed by List et

al. (2004), List (2007) and Levitt and List (2007) about whether real-money experiments really measure people's true preferences outside the experimental situation. Although it is arguably important to analyze various kinds of unselfish behavior experimentally, people may sometimes not be as unselfish or altruistic in a day-to-day setting as some experimental findings seem to suggest. Or they may indeed sometimes display remarkably unselfish behaviors also in real life, but these are often internally motivated and conditioned on the extent to which the individual herself receives credit for taking a certain action, not the actual consequences of this behavior per se.

References

Akerlof, G. A. and R. E. Kranton (2000). Economics and Identity, *Quarterly Journal of Economics*, 115(3), 715-53.

Akerlof, G and R. E. Kranton (2002). Identity and Schooling: Some Lessons for the Economics of Education, *Journal of Economic Literature*, 1167-1201.

Alpizar, F., Carlsson, F., and Johanson-Stenman, O. (2008), Anonymity, Reciprocity and Conformity: Evidence from Voluntary Contributions to a Natural Park in Costa Rica, *Journal of Public Economics*, 92, 1047-1060.

Andreoni, J. (1989). Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence, *Journal of Political Economy* 97(6), 1447-58.

Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving, *Economic Journal*, 100, 464-477.

Andreoni, J. (2006). Philanthropy, in *Handbook of the Economics of Giving, Reciprocity, and Altruism*, Volume 2, S.C, Kolm, and J.M. Ythier (Eds). Amsterdam: North-Holland.

Aquino, K., and A. Reed II. (2002). The Self-Importance of Moral Identity. *Journal of Personality and Social Psychology*, 83(6), 1423-1440.

Benabou, R., and J. Tirole. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3), 871-915.

Benabou, R., and J. Tirole. (2004). Willpower and personal rules. *Journal of Political* Economy, 112, 848-886.

Benabou, R., and J. Tirole. (2006). Incentives and prosocial behaviour. *American Economic Review*, 96(5), 1652-1678.

Baumeister, R., (1998). The self. In: Gilbert, D., Fiske, S., Lindzey, G. (Eds.), Handbook of Social Psychology. McGraw Hill, Boston.

Bennett, R. and R. Blaney (2002) Social consensus, moral intensity and willingness to pay to address a farm animal welfare issue. *Journal of Economic Psychology*, 23, 501-520.

Bernheim, D. B. (1994). A Theory of Conformity. *Journal of Political Economy*, 102 (5), 841-877.

Bodner, R. (1995). Self Knowledge and the Diagnostic Value of Actions: The Case of Donating to a Charitable Cause. Unpublished Ph.D. dissertation, MIT, Sloan School of Management.

Bodner, R., and D. Prelec. (2003). Self-signaling and Diagnostic Utility in Everyday Decision Making. *in* Brocas, I., and J. D. Carrillo (Eds). *The Psychology of Economic Decisions. Volume 1: Rationality and Well-Being*. Oxford: Oxford University Press.

Brekke, K. A., S. Kverndokk and K. Nyborg (2003), An economic model of moral motivation, *Journal of Public Economics*, 87, 1967-83.

Cameron, T. A., G. L. Poe, R. G. Ethier, and W. D. Schulze. (2002). Alternative Non-Market Value-Elicitation Methods: Are the Underlying Preferences the Same?, *Journal of Environmental Economics and Management*, 44, 391-425.

Carlsson F., Johansson-Stenman O. (2010). Scale factors and hypothetical referenda: A clarifying note, *Journal of Environmental Economics and Management*, 59, 286-292

Carlsson, F., and P. Martinsson. (2001). Do Hypothetical and Actual Willingness to Pay Differ in Choice Experiments? Application to the Valuation of the Environment, *Journal of Environmental Economics and Management*, 41, 179-92.

Carson, R. T. (1996), Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods, *Land Economics*, 72(1), 80-99.

Cherry, T. L., P. Frykblom and J. F. Shogren (2002), Hardnose the Dictator, *American Economic Review*, 92(4), 1218-21.

Cummings, R. G., S. Elliot, G. W. Harrison and E. E. Rutstrom (1995), Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible? *American Economic Review*, 85(1), 260-66.

Cummings, R. G., S. Elliot, G. W. Harrison, J. Murphy (1997). Are Hypothetical Referenda Incentive Compatible? *Journal of Political Economy*, 105(3), 609-21.

Cummings, R. and L. Taylor (1999). Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method, *American Economic Review*, 89, 649-665.

Forehand, M., R. Deshpandé, and A. Reed II. (2002). Identity Salience and the Influence of Differential Activation of the Social Self-Schema on Advertising Response. *Journal of Applied Psychology*, 87(6), 1086-1099.

Gilovich, Thomas, 1991, *Why We Know What Isn't So*, New York: The Free Press.

Harrison, G. W. and J. A. List (2004). Field Experiments. *Journal of Economic Literature* 42, 1009-1055.

Hensher, D, Louviere, J.J. and J. Swait (1999). Combining sources of preference data. *Journal of Econometrics* 89:197-221.

Herek, G. M. (1986). The Instrumentality of Attitudes: Toward a Neofunctional Theory. *Journal of Social Issues*, 42, 99-114.

Johansson-Stenman, O., and P. Martinsson. (2006). Honestly, why are you driving a BMW? *Journal of Economic Behavior and Organization*, 2006, 60, 129-46.

Johansson-Stenman, O., and H. Svedsäter. (2008). Measuring hypothetical bias in choice experiments: The role of cognitive consistency. *B-E. Journal of Economic Analysis and Policy*, 8, article 41.

Kahneman, D., and J. L. Knetsch. (1992). Valuing Public Goods: The Purchase of Moral Satisfaction, *Journal of Environmental Economics and Management*, 22, 57-70.

Kahneman, D., and Thaler, R. (2006), 'Anomalies: Utility Maximisation and Experienced Utility', *Journal of Economic Perspectives*, 20, 221-234.

Kahneman, D., Wakker, P., and Sarin, R. (1997), Back to Bentham? Explorations of Experienced Utility, *Quarterly Journal of Economics*, 112, 375-406.

Katz, D. (1960). The Functional Approach to the Study of Attitudes. *Public Opinion Quarterly*, 24, 163-204.

Kochi, I., B. Hubbel and R. Kramer (2003) *An empirical Bayes approach to combining estimates of the value of a statistical life for environmental policy analysis*, unpublished report to the U.S. Environmental Protection Agency.

Kuran, T. (1995). Private Truths, Public Lies: The Social Consequences of Preference Falsification (Harvard University Press, Harvard).

Lacetera, N., and M. Macis. (2010). Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme. *Journal of Economic Behavior and Organization*, 76, 225-237.

Levitt, S., and J. List. (2007). What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *Journal of Economic Perspectives*, 21(2), 153-174..

List, J. A., and C. A. Gallet. (2001). What Experimental Protocol Influence Disparities Between Actual and Hypothetical Values?, *Environmental and Resource Economics*, 20, 241-254.

List, J. A., P. Berrens, A. K. Bohara, and J. Kerkvliet (2004). Examining the Role of Social Isolation on Stated Preferences. *American Economic Review*. 94, 741-752.

List, J. A., P. Sinha, and M. H. Taylor. (2006). Using choice experiments to value non-market goods and services. *Advances in Economic Analysis & Policy*, 6(2), 1-37.

List, J. A. (2007). On the interpretation of giving in dictator games. Journal of Political Economy 115, 482–493.

Louviere, J.J, Hensher, D.A. and Swait, J.U.D. (2000). *Stated Choice Methods: Analysis and Applications*, Cambridge: Cambridge University Press.

Lusk, J. L., and T. C. Schroeder (2003), Are choice experiments incentive compatible? A test with quality differentiated beef steaks, *American Journal of Agricultural Economics*, 85, 840-56.

Murnighan, J. K., J. M. Oesch, and M. Pillutla. (2001). Player Types and Self-Impression Management in Dictatorship Games: Two Experiments, *Games and Economic Behavior*, 37, 388-414.

Murphy, J. J., P. G. Allen, T. H. Stevens, and D. Weatherhead. (2005). A Meta Analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics*, 30, 313-25.

Neilson, W.S., (2009). A theory of kindness reluctance, and shame for social preferences. Games and Economic Behavior 66, 394–403.

Nunes, P. and E. Schokkaert (2003). Identifying the warm glow effect in contingent valuation. *Journal of Environmental Economics and Management*, 45(2), 231–245.

Nyborg, K. and Brekke K. A (2010) Selfish Bakers, Caring Nurses? A Model of Work Motivation, *Journal of Economic Behavior and Organization* 75, 377-394.

Santos-Pinto, L and Sobel, J. (2005) A Model of Positive Self-Image in Subjective Assessments. *American Economic Review* 95, 1386-1402.

Shih, M., T. L. Pittinsky, and N. Ambady. (1999). Stereotype susceptibility, identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80-83.

Singer, P. (1974), All animals are equal, *Philosophical Exchange*, 1, 103-16.

Smith, A. (1759). *The Theory of Moral Sentiments*, Cambridge: Cambridge University Press.

Taylor, S. E., and J. D. Brown (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin* 116, 21-27.

Wardman, M. (2001). A review of British evidence on time and service quality valuations. *Transportation Research Part E*, 37, 107-128.

Appendix A: Proofs

Proof that Hypothesis 1 follows from the model

Let us start with the hypothetical treatment, implying that *x* and *WWF* are held constant. The maximization of *v* in eq. (4) then implies the maximization of *s*. The first order condition for an interior solution then implies

$$\frac{\partial f}{\partial d_{WWF}^{ethics}} \frac{\partial \left| MWTP_{WWF}^{stated} - MWTP_{WWF}^{moral} \right|}{\partial MWTP_{WWF}^{stated}} + \frac{\partial f}{\partial d_{WWF}^{honesty}} \frac{\partial \left| MWTP_{WWF}^{stated} - MWTP_{WWF}^{true} \right|}{\partial MWTP_{WWF}^{stated}} = 0 . \qquad \text{(A1)}$$

Since $\frac{\partial f}{\partial d_{WWF}^{ethics}} < 0$ and $\frac{\partial f}{\partial d_{WWF}^{honesty}} < 0$ we must have that one of the other two factors is negative.

Assume first that $\frac{\partial \left| MWTP_{WWF}^{stated} - MWTP_{WWF}^{moral} \right|}{\partial MWTP_{WWF}^{stated}} < 0$ and that $\frac{\partial \left| MWTP_{WWF}^{stated} - MWTP_{WWF}^{true} \right|}{\partial MWTP_{WWF}^{stated}} > 0$.

Then it follows that $MWTP_{WWF}^{stated} < MWTP_{WWF}^{moral}$ and that $MWTP_{WWF}^{stated} > MWTP_{WWF}^{true}$, which together implies that $MWTP_{WWF}^{true} < MWTP_{WWF}^{stated} < MWTP_{WWF}^{moral}$. Since in this treatment we have that $MWTP_{WWF}^{stated} = MWTP_{WWF}^{hyp}$, we have that

$$MWTP_{WWF}^{true} < MWTP_{WWF}^{hyp} < MWTP_{WWF}^{moral} . \qquad \text{(A2)}$$

Assume now instead that $\frac{\partial \left| MWTP_{WWF}^{stated} - MWTP_{WWF}^{moral} \right|}{\partial MWTP_{WWF}^{stated}} > 0$ and that $\frac{\partial \left| MWTP_{WWF}^{stated} - MWTP_{WWF}^{true} \right|}{\partial MWTP_{WWF}^{stated}} < 0$. Then it follows that $MWTP_{WWF}^{stated} > MWTP_{WWF}^{moral}$ and that $MWTP_{WWF}^{stated} < MWTP_{WWF}^{true}$. However, this implies that $MWTP_{WWF}^{true} > MWTP_{WWF}^{moral}$. This is a contradiction from the definition of a moral good, which implies that $MWTP_{WWF}^{true} < MWTP_{WWF}^{moral}$. Moreover, the shape of *f* makes the interior solution unique and rules out corner solutions where either $d_{WWF}^{ethics} = 0$ or $d_{WWF}^{honesty} = 0$. Thus (A2) holds.

Let us now turn to the real money treatment where the problem is to maximize (3), where *v* depends positively on both *U* and *s*. The choice experiment is constructed such that for a larger stated *MWTP* more *G* will be provided, implying less *x* for the subject. This implies that we can write

$$U = r(MWTP_{WWF}^{stated}) \quad , \tag{A3}$$

which has a maximum at $MWTP_{WWF}^{stated} = MWTP_{WWF}^{true} \equiv \frac{\partial u / \partial WWF}{\partial u / \partial Money}$. Substituting (A3) and (4) into (3) implies that we can write the first order condition for an interior optimum with respect to $MWTP_{WWF}^{stated}$ as

$$\frac{\mathrm{d}V}{\mathrm{d}MWTP_{WWF}^{stated}} = \frac{\partial v}{\partial U}\frac{\mathrm{d}r}{\mathrm{d}MWTP_{WWF}^{stated}} + \frac{\partial v}{\partial s}\frac{\mathrm{d}f}{\mathrm{d}MWTP_{WWF}^{stated}} = 0 \cdot \tag{A4}$$

We know from the above that *U* is maximized at $MWTP_{WWF}^{stated} = MWTP_{WWF}^{true}$, and that the maximum value of *s* is obtained at $MWTP_{WWF}^{stated} = MWTP_{WWF}^{hyp}$, where $MWTP_{WWF}^{hyp} > MWTP_{WWF}^{true}$. This implies that $\frac{\mathrm{d}V}{\mathrm{d}MWTP_{WWF}^{stated}} > 0$ at $MWTP_{WWF}^{stated} = MWTP_{WWF}^{true}$ and that $\frac{\mathrm{d}V}{\mathrm{d}MWTP_{WWF}^{stated}} < 0$ at $MWTP_{WWF}^{stated} = MWTP_{WWF}^{hyp}$. By the intermediate value theorem, this means that the value that maximizes *V*, $MWTP_{WWF}^{stated} = MWTP_{WWF}^{real}$, must fulfill

$$MWTP_{WWF}^{true} < MWTP_{WWF}^{real} < MWTP_{WWF}^{hyp} . \tag{A5}$$

Combining (A2) and (A5) together implies Hypothesis 1.

Proof that Hypothesis 2 follows from the model

Let us again start with the hypothetical treatment, implying that *x* and *Rest* are held constant. For the valuation of amoral goods we have by definition that $MWTP^{moral} = MWTP^{true}$. This

implies that s is maximized when $d_{Rest}^{ethics} = d_{Rest}^{honesty} = 0$ so that $MWTP_{Rest}^{stated} = MWTP_{Rest}^{true}$, i.e., the optimal response is to answer truthfully. Thus,

$$MWTP_{Rest}^{hyp} = MWTP_{Rest}^{true}. \quad \text{(A6)}$$

In the real money treatment we know from above that U is maximized at $MWTP_{Rest}^{stated} = MWTP_{Rest}^{true}$. Since U and s are maximized at the same $MWTP$ level, we have from (A4) that also V is maximized at this level, so that

$$MWTP_{Rest}^{Real} = MWTP_{Rest}^{true}. \quad \text{(A7)}$$

Combining (A6) and (A7) implies Hypothesis 2.

Appendix B: Experimental instructions

Introduction (all subjects)

On the following pages you will be asked to make a number of choices. Please read the instructions carefully before you make any decisions.

Please start with the first page and then continue to the second page and so on. Please do not go back and change your answer once you have proceeded to another page.

This is not a knowledge test—there are no right or wrong answers. We are interested in how people make decisions in these situations. Your answers will be treated anonymously. The only things we ask of you are your age and gender.

If you have any questions please do not hesitate to ask us.

Thank You!

Male ☐ Female ☐

Age: _____

Valuation task (all subjects)

This study is part of a research project aimed at assessing how much people value various goods and services. You will be asked both how much you value a visit to a local restaurant and how much you value a campaign run by the World Wildlife Fund (WWF).

The restaurant is an Italian restaurant called Il Gambero in Gothenburg (*address provided to subjects*). Il Gambero is a mid-priced restaurant that has received positive reviews in local newspapers. The WWF campaign was initiated to save the Asian Elephant, which is currently a mammal threatened by extinction. Among other things, the WWF is lobbying for the establishment of nature reserves for the elephants, and works to prevent various harmful activities jeopardizing the survival of the species.

In what follows you will be presented a number of choices. In each of these there are two alternatives to choose between, denoted Alternative A and Alternative B. Each alternative specifies how much money you will receive in cash and the value of a restaurant voucher valid at Il Gambero *or* how much will be donated to the WWF campaign. In other words, in some cases cash payments are set against a restaurant visit, and in other cases cash payments are set against a donation to the WWF fund. Please indicate which alternative you prefer in each choice situation. We furthermore want you to consider each choice independently. The total cash payment you will receive varies from SEK 0 to SEK 160, and the value of the voucher, or alternatively the donation to WWF, varies from SEK 0 to SEK 240.

Hypothetical scenario

Even if the choices below are hypothetical we would like you to carefully consider each of them as if they involved real monetary trade offs. Please note that we are interested in how much the particular restaurant visit and the WWF campaign are actually worth to you! It is hence important that you answer all questions truthfully. The restaurant voucher is valid for two months from today's date and may be used for you and any accompanying persons. Your stated donation to the WWF will be aggregated alongside donations made by other individuals, indicating the economic value participants in this study place on the campaign.

Your answers are treated anonymously. No one else but you will know how you have answered the questions.. Once you have completed the questionnaire you will be asked to put this in the empty envelope provided and seal this before handing it back to the research leader. He will then hand it to a secretary at the department who does not know anything about the purpose of the study. He or she will then enter your data into a data file that does not contain any identification details.

Real scenario (above text in italics replaced by what follows)

After you have made all choices, a draw will be made to decide which particular choice situation will determine your payoff. Depending on the alternatives of this choice you will in addition either be rewarded with a restaurant voucher, or a donation will be made to the WWF. Bear in mind that your choices will determine how much cash you may earn and the value of the restaurant voucher or the donation to the WWF campaign. It is hence important that you answer all questions truthfully.

Your answers are treated anonymously. No one else but you will know how you have answered the questions here. You will be given a note printed with a unique number. You must save this note; otherwise you will not receive any compensation! The same number is written on your questionnaire. When you have completed all choices, you will be asked to draw a number from 1 to 16 from a pot, specifying which choice situation will determine the actual cash payment made to you and the voucher value or donation. After the draw you may look at your answer for this particular choice situation.

Then you will be requested to put your questionnaire in the empty envelope provided and seal it. The number drawn from 1 to 16 will be written on the outside of this envelope before handing it back to the research leader. After this you are free to leave.

The envelope will later be handed to a secretary at the department who does now know anything about the purpose of the study. He or she will open the envelope, check which alternative you have chosen in the randomly drawn choice set, and put the corresponding cash payment in another envelope. In case the chosen alternative involves a restaurant voucher, this will also be put in the envelope. In case it involves a donation to the WWF, a receipt will be placed in the envelope confirming the total amount sent to the WWF by the research team, made as an anonymous donation. Please note that you will not receive a receipt of your personal contribution. Instead this will correspond to what all participants in this experiment have contributed together.

You will finally be notified via e-mail when and where you may collect your reward envelope.

Appendix C: Choice sets presented to subjects

Block A (half of respondents)

Choice #1	*Alt. A*	*Alt. B*
Money to you	80	0
Rest voucher	0	120

Choice #2	*Alt. A*	*Alt. B*
Money to you	0	120
Rest voucher	120	0

Choice #3	*Alt. A*	*Alt. B*
Money to you	40	80
Rest voucher	60	0

Choice #4	*Alt. A*	*Alt. B*
Money to you	120	40
Rest voucher	0	60

Choice #5	*Alt. A*	*Alt. B*
Money to you	160	40
Rest voucher	0	120

Choice #6	*Alt. A*	*Alt. B*
Money to you	40	80
Rest voucher	180	0

Choice #7	*Alt. A*	*Alt. B*
Money to you	40	120
Rest voucher	240	0

Choice #8	*Alt. A*	*Alt. B*
Money to you	160	40
Rest voucher	0	240

Choice #9	*Alt. A*	*Alt. B*
Money to you	0	40
WWF donation	120	0

Choice #10	*Alt. A*	*Alt. B*
Money to you	160	0
WWF donation	0	120

Choice #11	*Alt. A*	*Alt. B*
Money to you	40	160
WWF donation	60	0

Choice #12	*Alt. A*	*Alt. B*
Money to you	80	40
WWF donation	0	120

Choice #13	*Alt. A*	*Alt. B*
Money to you	40	120
WWF donation	120	0

Choice #14	*Alt. A*	*Alt. B*
Money to you	120	40
WWF donation	0	180

Choice #15	*Alt. A*	*Alt. B*
Money to you	40	160
WWF donation	180	0

Choice #16	*Alt. A*	*Alt. B*
Money to you	80	40
WWF donation	0	240

Block B (half of respondents)

Choice #17	*Alt. A*	*Alt. B*
Money to you	0	40
Rest voucher	120	0

Choice #18	*Alt. A*	*Alt. B*
Money to you	160	0
Rest voucher	0	120

Choice #19	*Alt. A*	*Alt. B*
Money to you	40	160
Rest voucher	60	0

Choice #20	*Alt. A*	*Alt. B*
Money to you	80	40
Rest voucher	0	120

Choice #21	*Alt. A*	*Alt. B*
Money to you	40	120
Rest voucher	120	0

Choice #22	*Alt. A*	*Alt. B*
Money to you	120	40
Rest voucher	0	180

Choice #23	*Alt. A*	*Alt. B*
Money to you	40	160
Rest voucher	180	0

Choice #24	*Alt. A*	*Alt. B*
Money to you	80	40
Rest voucher	0	240

Choice #25	*Alt. A*	*Alt. B*
Money to you	80	0
WWF donation	0	120

Choice #26	*Alt. A*	*Alt. B*
Money to you	0	120
WWF donation	120	0

Choice #27	*Alt. A*	*Alt. B*
Money to you	40	80
WWF donation	60	0

Choice #28	*Alt. A*	*Alt. B*
Money to you	120	40
WWF donation	0	60

Choice #29	*Alt. A*	*Alt. B*
Money to you	160	40
WWF donation	0	120

Choice #30	*Alt. A*	*Alt. B*
Money to you	40	80
WWF donation	180	0

Choice #31	*Alt. A*	*Alt. B*
Money to you	40	120
WWF donation	240	0

Choice #32	*Alt. A*	*Alt. B*
Money to you	160	40
WWF donation	0	240

Figure 1. Examples of choice sets used

	Alternative A	*Alternative B*
Money that you will receive	40	120
Voucher at local restaurant	120	0

	Alternative A	*Alternative B*
Money that you will receive	160	40
WWF donation	0	240

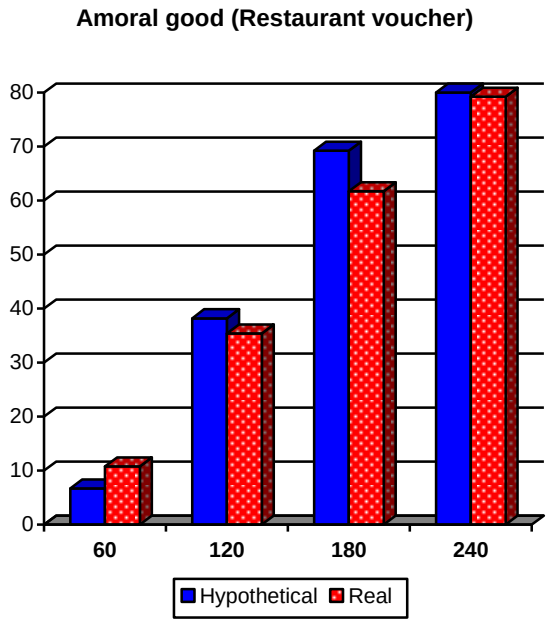

Figure 2. Relative frequency of choices of the amoral good (restaurant voucher) versus cash payment.
The horizontal axis displays the size of the voucher, all else equal.

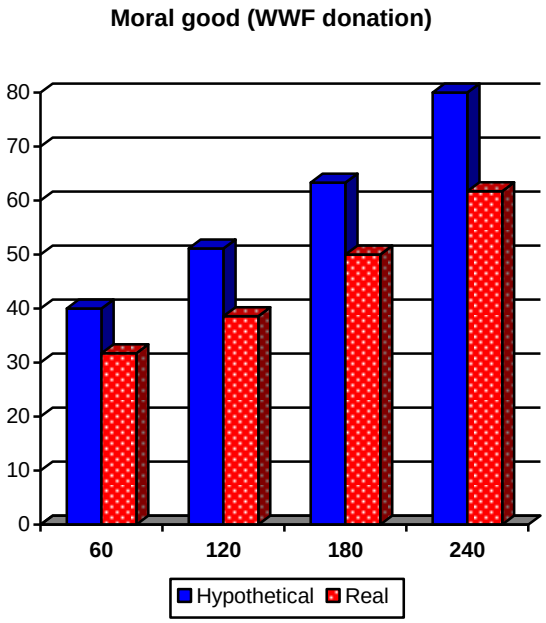

Figure 3. Relative frequency of choices of the moral good (WWF donation) versus cash payment.
The horizontal axis displays the size of the donation, all else equal.

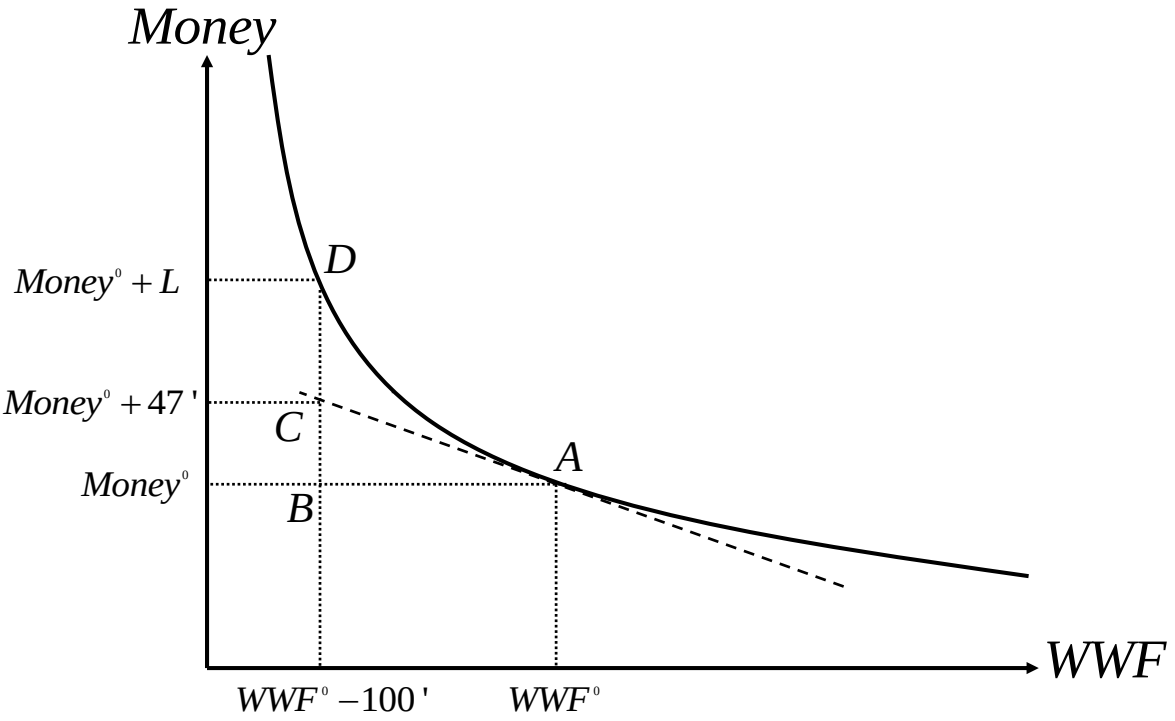

Figure 4. Indifference curve in *Money*-*WWF*-space.

Table 1. Estimated parameters from pooled regression models (standard errors in parentheses).

	Logit, clustered, robust standard errors	Random effects logit	Fixed effects logit
Money to subjects, β	0.027*** (0.0016)	0.035*** (0.0016)	0.035*** (0.0018)
Restaurant voucher money, ρ	0.014*** (0.0013)	0.019*** (0.0012)	0.019*** (0.0015)
WWF donation money, γ	0.013*** (0.0012)	0.017*** (0.0012)	0.017*** (0.0015)
Hypothetical treatment times restaurant voucher, λ	0.0014 (0.0017)	0.002 (0.0015)	0.002 (0.0021)
Hypothetical treatment times WWF donation, μ	0.005*** (.0009)	0.007*** (0.0015)	0.007*** (0.0022)
Log-likelihood function	-1367.686	-1203.705	-781.594
Statistical observations	2560	2560	2560
Number of subjects	160	160	160

*Note:****, **, and * denote significance at the 0.01 level, 0.05 level, and 0.1 level, respectively.

Table 2. Estimated *MWTPs* and *MWTP*-based measures of hypothetical bias (standard errors calculated using the delta method in parentheses).

	logit, clustered, robust standard errors	Random effects logit	Fixed effects logit
	Baseline real-money marginal willingness to pay estimates		
$MWTP_{Rest}^{Actual}$	0.540*** (0.038)	0.539*** (0.030)	0.545*** (0.045)
$MWTP_{WWF}^{Actual}$	0.474*** (0.043)	0.471*** (0.029)	0.474*** (0.043)
	Measures of hypothetical bias		
$MWTP_{Rest}^{Hyp} - MWTP_{Rest}^{Actual}$	0.053 (0.052)	0.058 (0.042)	0.064 (0.061)
$MWTP_{WWF}^{Hyp} - MWTP_{WWF}^{Actual}$	0.184*** (0.063)	0.188*** (0.043)	0.193*** (0.061)

*Note:****, **, and * denote significance at the 0.01 level, 0.05 level, and 0.1 level, respectively.