

Article

A Structural Theory of the Willingness-to-Pay Gap: Valuation and Behaviour in Climate Mitigation

Haochen Ni¹

¹ Student (BSc Economics, 2024-2027), Department of Economics, University College London, United Kingdom; zctphni@ucl.ac.uk

How to cite

Ni, H. (2026). A Structural Theory of the Willingness-to-Pay Gap: Valuation and Behaviour in Climate Mitigation. *UCL Journal of Economics*, vol. 5 no. 1. DOI: 10.14324/111.444.2755-0877.2238

Peer review

This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymised during review.

Copyright

2026, *Haochen Ni*. This is an open-access article distributed under the terms of the Creative Commons Attribution Licence (CC BY) 4.0 <https://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited • DOI: 10.14324/111.444.2755-0877.2238

Open access

UCL Journal of Economics is a peer-reviewed open-access journal

Abstract

This paper offers a structural, micro-founded interpretation of the divergence between stated and behavioural willingness to pay (WTP) for climate mitigation. While the value-action gap is commonly attributed to hypothetical bias or behavioural distortions, we show that it can arise endogenously even under fully rational behaviour. The key insight is that welfare-based WTP, defined as compensating variation, is characterised by a level condition on utility, whereas behavioural WTP is determined by marginal conditions within an optimisation problem. Embedding both objects within a unified framework, we show that a wedge between the two measures arises generically, except under knife-edge conditions. We further introduce equilibrium feedback to capture the hypothetical nature of survey environments and analyse the role of observability in shaping behavioural incentives. Finally, we discuss SCC models, cheap talk, and empirical testing as extensions and implications of the framework.

Keywords: *Willingness to pay, Compensating variation, Behavioural optimisation, Public goods, Environmental economics, Social image, Value-action gap*

1. Introduction

Imagine being asked to complete a survey about how much one is willing to pay for climate change mitigation. How much would one answer? What factors would one consider when answering the question? Would one really contribute as much as one answered? A common challenge in environmental economics is that individuals' stated willingness to pay (WTP) for climate mitigation often exceeds their actual contributions, a phenomenon known as the "value-action gap". This divergence is typically attributed to hypothetical bias and moral considerations.

This paper offers a different perspective. We argue that the gap arises because stated valuation WTP is defined by a level condition, whereas behavioural contribution is governed by marginal incentives. They correspond to fundamentally different economic concepts. In standard welfare analysis, WTP is defined as compensating variation (CV), a money-metric measure derived from utility comparisons across states. By contrast, voluntary contributions are determined by individual optimisation and reflect marginal incentives at the optimum. As a result, even under fully rational behaviour and consistent preferences, the two measures need not coincide.

This distinction also provides a structural interpretation of survey-design interventions such as cheap talk design (Cummings and Taylor, 1999; Murphy, Stevens and Weatherhead, 2005). In particular, it aims to align hypothetical responses with actual behavioural willingness to pay by reminding respondents of budget constraints and opportunity costs. While cheap talk is commonly viewed as a method to reduce bias, our framework suggests that it may also operate by shifting respondents from a welfare-valuation frame towards a behavioural choice frame. In other words, the effectiveness of cheap talk may partly arise because it changes the economic object being elicited.

To formalise this distinction, we develop a unified framework in which both welfare-based valuation and behavioural contribution are defined within the same preference structure. We show that a wedge between the two arises generically, except under knife-edge conditions (i.e., special cases requiring additional restrictions on the model primitives). We further consider equilibrium feedback to capture the hypothetical environment in the survey and to extend the model accordingly. We also examine the role of observability of pro-environmental actions as a mechanism linking valuation and behaviour. When contributions are visible, individuals obtain reputational benefits that strengthen marginal incentives. Because behavioural responses depend on marginal incentives while welfare-based valuation reflects state comparisons, observability can affect the two measures differently and potentially reduce the wedge.

The remainder of the paper is organised as follows. Section 2 discusses the related literature. Sections 3.1-3.3 develop a static unified framework, and Section 3.4 establishes the divergence result. Section 3.5 introduces the equilibrium and suggests that observability can serve as an incentive shift, helping align behaviour with valuation. Section 3.6 attributes specific functional forms and substitutes numerical values for illustration. Section 3.7 proposes an executable test for observability to reduce the gap, and Section 4 concludes. Section 5 discusses potential extensions.

2. Literature Review

Willingness to pay (WTP) is traditionally grounded in Hicksian welfare theory, where it is defined as compensating variation (Hicks, 1939; Hanemann, 1984). Under this framework, WTP is a normative object that reflects the income adjustment required to equate utility before and after an environmental change. Subsequent work has established conditions under which stated-preference methods can recover welfare-consistent measures, such as the consequential design of a survey (Carson and Groves, 2007), reinforcing the interpretation of WTP as a welfare-based, state-dependent valuation.

A separate strand of literature models environmental contributions as the outcome of individual optimisation. In Andreoni's (1990) impure altruism framework, individuals derive utility both from the level of the public good and from the act of giving itself. Behavioural contributions are therefore determined by marginal incentives and shaped by factors such as social image and reputational concerns (Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009). Experimental evidence, such as in Ariely, Bracha, and Meier (2009) and in DellaVigna et al. (2012), shows that observability and social expectations significantly affect contribution behaviour. In this literature, WTP is not a welfare measure but an endogenous choice variable.

The divergence between stated valuation and observed behaviour is typically attributed to hypothetical bias, moral expression, or informational frictions (Kahneman and Knetsch, 1992; List and Gallet, 2001; Alpizar, Carlsson, and Johansson-Stenman, 2008). Carson (2012), by contrast, defends contingent valuation methods by arguing that, under appropriate survey design, stated responses can support valid welfare analysis. Despite these differences, both perspectives emphasise measurement issues

or behavioural distortions; they implicitly treat stated WTP and realised contributions as imperfect measures of the same underlying construct.

Nevertheless, two papers explicitly argue that not every revealed valuation gap should be dismissed as bias or measurement error. Zhao and Kling (2004) question the equivalence between WTP and Hicksian welfare measures in dynamic settings with uncertainty and commitment costs. Our equilibrium feedback introduced in Section 3.5 and the dynamic extensions outlined in Section 5 are closer in spirit to Zhao and Kling, since both approaches make beliefs and timing central. However, the mechanisms differ. Zhao and Kling focus on private learning under uncertainty, whereas our extension emphasises social learning through observability and beliefs about the environmental state. Bartling, Engl, and Weber (2015) show that WTP-WTA (willingness-to-accept) gaps persist even among subjects who understand the incentive properties of the Becker-DeGroot-Marschak mechanism, suggesting that such gaps need not be driven solely by game-form misconceptions. Our focus, on the other hand, is to analyse the structural divergence between welfare-based WTP and behavioural contribution.

At the aggregate level, climate valuation is often benchmarked against the social cost of carbon (SCC), which is derived as the shadow price of emissions within an intertemporal welfare maximisation framework (Nordhaus, 2017). Although the SCC is defined by a marginal condition in the social planner's problem, it remains a normative welfare metric used for cost-benefit analysis rather than a behavioural contribution chosen by individuals. In this sense, SCC-based benchmarks are closer in purpose to welfare-based valuation than to voluntary contribution behaviour. Consequently, empirical comparisons between individual WTP and SCC-based benchmarks are often interpreted as evidence of under- or over-valuation, without recognising that these comparisons may involve fundamentally different economic objects.

A broader theoretical foundation for this distinction is found in the literature on revealed preference, behavioural welfare, and prosocial motivation. Sen's (1977) critique of revealed preference argues that observed choices should not be mechanically interpreted as welfare-maximising preferences, since behaviour may also reflect commitment, sympathy, or moral reasoning. Similarly, Bernheim and Rangel (2009) show that welfare analysis becomes non-trivial when choices depend on ancillary conditions of the decision environment, such as framing or contextual features that affect which preference is revealed. In methodological terms, this is an identification problem: as Manski (2004) emphasises, observed choice generally reflects preferences, beliefs, and other factors, making it difficult to infer the underlying decision process from behavioural data alone. These contributions suggest that a divergence between stated valuation and realised contribution should not be interpreted solely as bias or inconsistency; rather, the two may arise from distinct mechanisms activated by different elicitation environments.

Following this literature, we also do not interpret stated and behavioural WTP solely as noisy measures of a single construct; instead, we argue that the divergence can arise structurally because the two measures correspond to fundamentally different economic objects. Compensating variation is defined by a level condition on utility across states, whereas behavioural contributions are determined by marginal conditions within an optimisation problem. By modelling both within a unified framework and incorporating observability as a driver of marginal incentives, we show that the value-action gap can emerge endogenously even in the absence of bias or irrationality.

Finally, as we mentioned in the introduction, an emerging approach to reduce hypothetical bias and improve the validity of stated preference surveys is the "cheap talk" design (Cummings and Taylor, 1999; Murphy, Stevens and Weatherhead, 2005), which aims to align hypothetical responses with actual behavioural willingness to pay by reminding respondents of budget constraints and opportunity costs. This is directly relevant to the present framework. In this paper, we argue that one reason the value-action gap exists is that standard contingent-valuation surveys are typically designed to elicit a welfare-based WTP measure close to the compensating variation, whereas actual contribution behaviour is governed by marginal incentives. Therefore, when researchers want stated responses to predict actual contributions rather than welfare valuation, survey design may need to shift respondents from a level-comparison frame towards a behavioural-choice frame. Cheap talk scripts may partly serve this role by emphasising budget constraints and foregone consumption. They can prompt respondents to consider the marginal cost of giving, leading to WTP estimates closer to the behavioural contribution. On this interpretation, cheap talk may reduce not only hypothetical bias but also help align the elicited welfare measure with the behavioural object that researchers intend to predict.

3. The Model

3.1. Primitives and Assumptions

We start by assuming the utility depends on private consumption that generates emissions c ; willingness to pay $W \in [0, c]$; environmental quality $E(\cdot)$; and a moral component $M(\cdot)$. Hence, the utility function for an individual i is:

$$U_i = u_i(c - W_i) + E_i(\cdot) + M_i(\cdot)$$

We impose a regularity condition to allow subsequent analysis.

A1 (Regularity) *The general utility function is additively separable and twice continuously differentiable.*

We further define environmental improvement as an aggregate outcome resulting from both public policy and voluntary contributions. Let $Q \geq 0$ denote the actual improvement resulting from total private giving ($\sum_j W_j$) and public policy (τ):

$$Q = \tau + W_i + \sum_{j \neq i} W_j$$

Individuals might not fully perceive the environmental improvement. Therefore, we introduce a parameter $p_i \in [0, 1]$, representing the proportion of the environmental improvement that the individual experiences. Consequently, the subjective environmental improvement relevant to the individual i is modelled as:

$$E_i = E(p_i Q)$$

where $E(\cdot)$ is a transformation function converting environmental quality into utility.

For moral satisfaction, following Andreoni (1990), we assume moral satisfaction depends on a combination of the warm-glow effect (M_i^{self}) and a social-image component related to personal reputational payoff (M_i^{social}), so we write M_i as:

$$M_i(W_i) = M_i^{self}(W_i) + M_i^{social}(W_i, s_i, \lambda_i)$$

In particular, social image is modelled as an expected value that depends on the observability of one's pro-environmental action $s_i \in [0, 1]$ and the reputational return $\lambda_i \geq 0$.

A2 (Monotonicity and concavity) *All component utility functions are monotonic and concave. In particular,*

$$\begin{aligned} u'(c - W_i) &> 0, & u''(c - W_i) &< 0, \\ E'(p_i Q) &\geq 0, & E''(p_i Q) &\leq 0, \\ M'(W_i) &\geq 0, & M''(W_i) &\leq 0. \end{aligned}$$

We derive Lemma 1 directly from A2.

Lemma 1 (General utility concavity) *Under A1 and A2, U_i is strictly concave in W_i implying a unique interior optimum whenever it exists.*

Remark 1 (Scaling of observability) For $s_i \in [0, 1]$, we interpret observability as the degree of visibility, normalised so that $s_i = 0$ denotes complete anonymity and $s_i = 1$ denotes full public disclosure. If observability were measured on a different scale, say $s' \in [0, S_{\max}]$, all results would scale accordingly. Although the derivative magnitudes may vary, signs are preserved.

3.2. WTP as Compensating Variation

Start with the compensating variation framework. Consider a change in the economic environment from state 0 to 1, the compensating variation W_i^{CV} for the individual i is defined implicitly by:

$$u(c - W_i^{CV}) + E(p_{i,1}Q_1(W_i^{CV})) + M(W_i^{CV}) = u(c) + E(p_{i,0}Q_0(0)) + M(0)$$

where we assumed that, in the baseline period, the individual contributed nothing. This defines WTP as the income adjustment required to hold utility constant across states, i.e. as a level condition on utility rather than a marginal condition.

Within this compensating variation framework, WTP is treated as a money-metric measure of the change in welfare between two states. WTP is shown to be an implied valuation derived from welfare comparisons, not the result of individual utility maximisation. In the next section, we will consider the case in which individuals actively choose their WTP under certain behavioural assumptions.

3.3. WTP as Optimisation Behaviour

“When people make donations to privately provided public goods, they may not only gain utility from increasing its total supply, but they may also gain utility from the act of giving.”

--- J. Andreoni (1990, *The Economic Journal*)

This quote motivates my approach to modelling WTP as an optimisation problem, and hence, I discovered the divergence between frameworks. If an individual derives utility from the act of giving itself, then WTP is no longer just a descriptive measure of welfare change. It becomes an active choice.

To frame WTP as the outcome of individual optimisation, we need to revise a few assumptions about variables. We write

$$E_i = E(p_i \left(\tau + W_i + \sum_{j \neq i} W_j \right))$$

Assumption A2 still holds. We also model moral satisfaction as a result of the choice W_i :

$$M(W_i) = M^{self}(W_i) + M^{social}(W_i, s_i, \lambda_i)$$

Now we can frame WTP as a utility maximisation problem:

$$W_i^* \in \arg \max_{W_i \in [0, c]} u(c - W_i) + E(p_i \left(\tau + W_i + \sum_{j \neq i} W_j \right)) + M^{self}(W_i) + M^{social}(W_i, s_i, \lambda_i)$$

We will not assign specific functional forms to this optimisation problem for now to derive an exact answer, but we will sketch the condition for an interior solution to exist.

Remark 2 (Interior solution) Under Lemma 1, the utility function $U(W_i)$ is strictly concave. Suppose $U(W_i)$ is continuously differentiable on $(0, c)$. If

$$\lim_{W \rightarrow 0^+} U'(W) > 0 \quad \text{and} \quad \lim_{W \rightarrow c^-} U'(W) < 0$$

then there exists a unique interior solution $W_i^* \in (0, c)$.

Therefore, the FOC is

$$U'(W_i^*) = -u'(c - W_i^*) + p_i E'(p_i Q(W_i^*)) Q'(W_i^*) + M'(W_i^*) = 0$$

where $Q = \tau + W_i^* + \sum_{j \neq i} W_j$ and all functions are the derivatives with respect to W_i .

In Remark 2, we mentioned the marginal utility $U'(W)$ at two limits, 0 and c . Intuitively, the condition $\lim_{W \rightarrow c^-} U'(W) < 0$ is likely satisfied, since when W tends to c , consumption tends to zero, and hence by the standard result of diminishing marginal utility, $u'(c - W_i)$ should be high, and hence making $U'(W_i)$ negative. On the other hand, whether $\lim_{W \rightarrow 0^+} U'(W) > 0$ is not guaranteed. If the condition fails, then the corner solution $W_i = 0$ is obtained, meaning the individual is unwilling to contribute anything.

3.4. The Wedge Between Two WTPs

To see how the two frameworks diverge, we define the wedge as the welfare-based valuation minus the behavioural contribution, which is also the value-action gap. Explicitly, we write

$$\Delta_i = W_i^{CV} - W_i^*$$

where W_i^{CV} is the WTP derived from compensating variation and W_i^* derived from optimisation. It can be easily seen that the wedge exists because the two frameworks model fundamentally different economic concepts; its existence is formally stated below.

Proposition 1 (Existence of the wedge) *Under A1 and A2, and assuming an interior solution exists, W_i^* and W_i^{CV} are generally distinct, so $\Delta_i \neq 0$ except for knife-edge conditions, when the level condition and the first-order condition coincide.*

Proof. Let U_i^0 denote utility in the initial state, which is purely exogenous. Then, the benchmark compensating variation W_i^{CV} is defined by

$$U(W_i^{CV}) = U_i^0. \quad (1)$$

The behavioural optimum W_i^* is determined by

$$U_W(W_i^*) = 0. \quad (2)$$

By strict concavity (A2), $U_W(W_i) = 0$ has a unique solution W_i^* and $U(W_i) = U_i^0$ has at most one solution W_i^{CV} . Assume, for contradiction, that

$$W_i^* = W_i^{CV} = \tilde{W}.$$

Then, $\tilde{W}$ must satisfy both equations (1) and (2):

$$\begin{aligned} U(\tilde{W}) &= U_i^0 \\ U_W(\tilde{W}) &= 0, \end{aligned}$$

which implies

$$U_W(W_i^{CV}) = 0. \quad (3)$$

Equation (3) implies that the derivative is zero on a given level set, which is an additional scalar restriction on the primitives. Assumptions A1-A2 do not imply this restriction. Hence, except for knife-edge conditions, generically

$$W_i^* \neq W_i^{CV}.$$

Q.E.D.

3.5. Introducing Equilibrium Feedback in CV

The above sections allowed a direct comparison between compensating variation and behaviour optimisation, and Section 3.4 demonstrates the structural divergence between the two frameworks. In this section, we propose that a better way to model it is to add equilibrium feedback in order to capture the real survey environment.

Again, consider a change in the economic environment from state 0 to 1. We distinguish between the baseline WTP, W_i^{CV} and denote equilibrium-based WTP, W_i^E . The compensating variation W_i^E for the individual i is defined implicitly by:

$$u(c - W_i^E) + E(p_{i,1}Q_1(\bar{W}_0)) + M(\bar{W}_1) = u(c) + E(p_{i,0}Q_0(\bar{W}_0)) + M(\bar{W}_0)$$

where we assume moral satisfaction and environmental quality in period t depends on the equilibrium voluntary contribution $\bar{W}_t$, rather than an individual's compensating variation payment W_i^E . We interpret $\bar{W}_t$ as a perceived equilibrium benchmark, which is deduced from the beliefs about the equilibrium and past observations. Rather than solving for a full strategic equilibrium, we adopt a reduced-form approach in which individuals form beliefs about aggregate behaviour based on past observations. In particular, we model $\bar{W}_t$ as a state variable capturing these beliefs, which may be anchored to past realised contributions, e.g. $\bar{W}_t = W_{t-1}^E$. This captures the idea that survey responses are evaluated in a hypothetical equilibrium environment rather than a purely exogenous state. For the purposes of the static analysis, we treat $\bar{W}_t$ as exogenous. Similarly, define the wedge as:

$$\Delta_i = W_i^E - W_i^*$$

It is natural to imagine that the wedge still exists and, therefore, Proposition 1 is still true. To avoid confusion, we write it as Proposition 1.1, and the proof is moved to Appendix A.

Proposition 1.1 (Existence of the wedge under equilibrium) Under A1 and A2, and assuming an interior solution exists, W_i^* and W_i^E are generally distinct, so $\Delta_i \neq 0$ except for knife-edge conditions, when the level condition and the first-order condition coincide.

Now, we want to analyse this result further. Note that one variable we introduced in both WTP frameworks is the observability, s . Since the compensating variation framework implies a level condition and the behavioural one implies marginal incentives, we speculate that the observability generates different magnitudes of effect. To prove this, we try to evaluate s on W_i^* and W_i^E . The behavioural framework in its general form makes it hard to directly determine the sign, but we can instead evaluate locally around the optimum, $W_i^*(s)$.

Lemma 2 (Sign of the wedge) Under Lemma 1, for any $w \in [0, c]$,

$$w \leq W_i^*(s) \iff U_W(w; s) \geq 0.$$

Since we define the wedge Δ_i as $\Delta_i = W_i^E(s) - W_i^*(s)$. Therefore,

$$\Delta_i \leq 0 \iff U_W(W_i^E(s); s) \geq 0.$$

The proof is in Appendix A. Lemma 2 allows us to evaluate the sign of the wedge without explicitly solving for W_i^* , by instead evaluating marginal incentives at the compensating variation benchmark W_i^E .

Remark 3 (Sign of the wedge under weak marginal incentives) Suppose at a low level of observability ($s \rightarrow 0$),

$$U_W(W_i^E; s) < 0$$

Then, under Lemma 2,

$$\Delta_i > 0$$

Remark 3 considers a situation in which observability is sufficiently low so that the marginal reputational return from contributing is weak. In this case, the marginal benefit from contributing is relatively small compared with the marginal utility cost of reducing private consumption. If the cost, in fact, dominates the variation in utility, then the behavioural marginal utility evaluated at W_i^E is negative. This condition is likely when the individual environmental impact p_i is low, and reputational incentives $\lambda_i s_i$ are weak. An anonymous survey environment closely corresponds to this, in which individual actions have low environmental effects and social recognition. Similarly, when observability is sufficiently high ($s \rightarrow 1$), reputational

incentives may dominate the private marginal cost of contributing, thereby creating a negative wedge, i.e., actual contributions exceed stated WTP.

After evaluating the general sign conditions for when s varies, we want to next show whether a shift in s indeed makes the wedge shrink. We show this by differentiating the wedge with respect to observability and deriving the necessary condition in Proposition 2.

Proposition 2 (Sensitivity condition) Since $\Delta_i = W_i^E(s) - W_i^*(s)$, we have

$$\frac{d\Delta_i}{ds} < 0 \iff \frac{dW^*}{ds} > \frac{dW_i^E}{ds}$$

Proof. The proof is rather straightforward. Suppose $W_i^E(s)$ and $W^*(s)$ are differentiable in s . Differentiating gives

$$\frac{d\Delta_i}{ds} = \frac{dW_i^E}{ds} - \frac{dW^*}{ds}$$

Therefore, $(d\Delta_i)/ds < 0$ if and only if

$$\frac{dW^*}{ds} > \frac{dW_i^E}{ds}$$

Hence, the wedge shrinks whenever behavioural WTP responds more strongly to observability than compensating-variation WTP.

Q.E.D.

Proposition 2 is not intended as a deep theorem by itself. It translates the theoretical distinction into a testable sensitivity condition, which we explain further in Section 3.7, where an empirical test of observability's effectiveness is sketched.

To evaluate the local scenarios, we first determine the continuity of WTP values in s , and the crossing conditions around the optimal observability. The proofs are moved to Appendix A.

Lemma 3 (Continuity in s) Under A1 and A2, U is twice continuously differentiable and strictly concave in W . Hence, by the Implicit Function Theorem, the interior solution $W^*(s)$ is continuous in s . Similarly, the compensating variation equation defines $W^E(s)$ as the unique solution to a strictly monotone equation in W , implying continuity in s .

Corollary 1 (Crossing condition) Suppose $W_i^*(s)$ and $W_i^E(s)$ are continuous in s . If

$$U_W(W_i^E(0); 0) < 0 \quad \text{and} \quad U_W(W_i^E(1); 1) > 0$$

Then, under Lemma 2, the conditions imply $\Delta_i(0) > 0$ and $\Delta_i(1) < 0$. By the Intermediate Value Theorem,

$$\exists s^* \in (0,1) \text{ such that } \Delta_i(s^*) = 0$$

If, additionally, $\Delta'_i(s^*) \neq 0$, the crossing is locally unique and sign-changing.

Finally, we state the local scenarios around the optimal s^* using Corollary 1.

Proposition 3 (Local valuation-consistency and observability) Assume $\Delta_i(s)$ is continuous and $\Delta'_i(s^*) \neq 0$. If Corollary 1 holds, then there exists $\varepsilon > 0$ such that

$$\Delta_i(s) > 0 \text{ for } s \in (s^* - \varepsilon, s^*) \text{ and } \Delta_i(s) < 0 \text{ for } s \in (s^*, s^* + \varepsilon)$$

Then, three cases arise locally around any such s^* : (i) Under-contribution, $s < s^*$; (ii) Valuation-consistency, $s = s^*$; (iii) Over-contribution, $s > s^*$.

In particular, (i) states that under-contribution occurs because the marginal social incentive is insufficient to overcome the private cost of contributing. (ii) indicates that when observability is at a valuation-consistent level, the individual contributes exactly as much as they value the environment. It suggests that behavioural contribution aligns with individual welfare valuation under the compensating variation framework. Finally, (iii) states that a sufficiently high s implies that the individual is “forced” by social stakes to spend more than they actually value the environment. This results in over-contribution. Figure 1 below shows a case in which an increase in observability s can align the two WTP values.

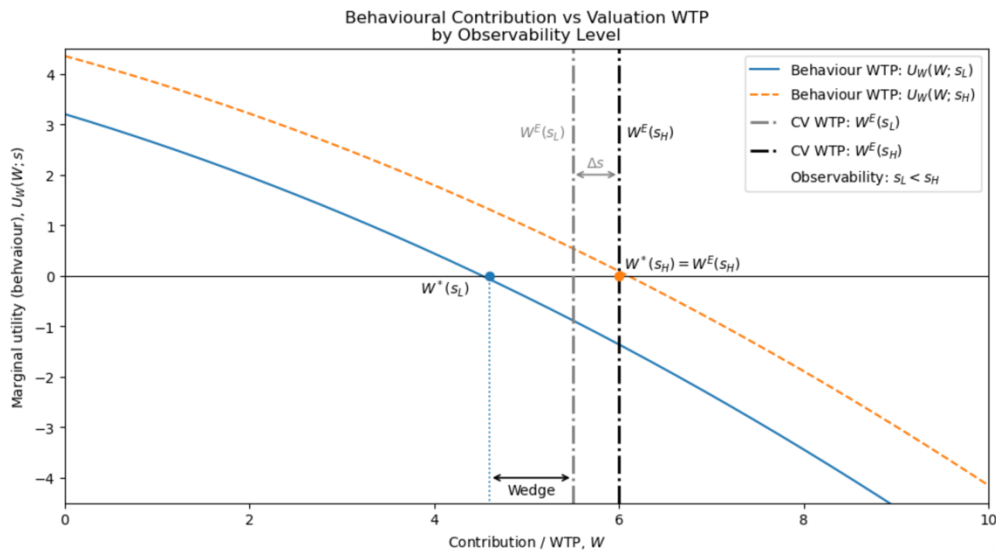


Figure 1: Behavioural contribution and welfare-based valuation under different levels of observability.

Notes: Welfare-based WTP $W^E(s)$ is determined by compensating variation and corresponds to a level condition. The behavioural contribution $W^*(s)$ is defined by the marginal condition $U_W(W; s) = 0$. They need not coincide. As illustrated, an increase in observability shifts behavioural incentives through social-image effects, potentially aligning behavioural contributions with welfare-based valuation at high observability s_H .

Remark 4 (Regarding the Social Cost of Carbon (SCC) model) If a planner sets policy based on W^{CV} (the normative value), but individuals act according to W^* (the positive behaviour), the resulting environmental change, Q , may be sub-optimal unless $s = s^*$. This suggests that “nudging” through observability can serve as a mechanism for aligning behavioural incentives with welfare-based valuation.

Remark 5 (Regarding cheap talk as object alignment) After formally establishing our model, we want to briefly discuss the model of cheap talk, which was briefly discussed in Section 2. The main model of this paper is not a model of cheap talk itself. Rather, it clarifies one channel through which cheap talk may be effective in stated-preference settings. If standard CV questions elicit a welfare-based object while actual behaviour is governed by marginal incentives, then cheap talk scripts that emphasise budget constraints and foregone consumption may reduce the wedge by shifting respondents toward the behavioural object. Thus, cheap talk can be interpreted as a partial object-alignment device rather than merely a generic correction for hypothetical bias.

3.6. A Parametric Example

Let the individual’s utility take the specific functional forms satisfying assumptions A1 and A2:

$$U(W) = \ln(c - W) + \ln(1 + p(\tau + NW)) + \ln(1 + W) + \lambda s W,$$

Table 1: Parametric Example Results

s	$W^E(s)$	$W^*(s)$	Wedge $\Delta_i = W_i^E(s) - W_i^*(s)$
0.0	90.91	66.48	+24.42
0.2	90.91	95.47	-4.56
0.5	90.91	98.08	-7.17
0.8	90.91	98.78	-7.87
1.0	90.91	99.02	-8.11

Notes: The table reports welfare-based WTP and behavioural WTP under different levels of observability. Parameter values are fixed at $c = 100$, $p = 0.1$, $N = 100$, $\tau = 0$, $\tau' = 100$, $\lambda = 1$, Observability varies over $s \in [0,1]$.

where the first term on the right-hand side is the private consumption, followed by environmental quality, warm-glow and social image. We multiplied W by N in the environmental quality function to represent symmetrical individuals with N individuals in total.

The individual obtains her voluntary contribution by maximising $U(W)$ with respect to $W \in (0, c)$. The first-order condition is given by

$$U'(W^*; s) = -\frac{1}{c - W^*} + \frac{pN}{1 + p(\tau + NW^*)} + \frac{1}{1 + W^*} + \lambda s = 0.$$

By contrast, the WTP under the compensating variation framework with equilibrium feedback is defined by the level condition:

$$\begin{aligned} \ln(c - W^E) + \ln(1 + p(\tau' + N\bar{W})) + \ln(1 + \bar{W}) + \lambda s \bar{W} \\ = \ln(c) + \ln(1 + p(\tau + N\bar{W})) + \ln(1 + \bar{W}) + \lambda s \bar{W} \end{aligned}$$

where we assumed the only exogenous change between the two states is a change in the government expenditure from τ to τ' . (The assumption is set to make both sides not cancel each other out completely.) We further isolate the observability channel in behavioural incentives by setting the equilibrium WTP $\bar{W}$ to be zero. Then,

$$\ln(1 + \bar{W}) + \lambda s \bar{W} = \ln(1) + 0 = 0,$$

and

$$W^E = c \left(1 - \frac{1 + p\tau}{1 + p\tau'} \right).$$

i.e. W^E is independent of s , allowing for a clearer comparison between the two frameworks.

Solving for both WTPs completely from the equations can be tedious and obscure. Therefore, to illustrate the mechanism, we assign numerical values to the parameters and examine how variations in observability affect the wedge.

Table 1 above shows that for low observability ($s = 0.0$), behavioural WTP ($W^* = 66.48$) is well below the constant compensating variation WTP ($W^E = 90.91$), indicating under-contribution. After the intersection point at $s^* \approx 0.09$, behavioural WTP exceeds W^E , leading to over-contribution. As observability increases further, the wedge becomes increasingly negative, indicating over-contribution. This illustrates the local regimes in Proposition 3: under-contribution at low observability and over-contribution at high observability.

3.7. Empirical Test for Observability

In Proposition 2, we compute the derivatives of WTP with respect to the observability to test for the response sensitivity. However, under the general functional form we used in this paper, it is ambiguous whether observability reduces the wedge. In Section 3.6, we assigned specific functional forms and numerical values to illustrate how the mechanisms work in practice. However, we have made a few assumptions, and they do not necessarily reflect reality. Hence, in this section, we will sketch a testable empirical design for Proposition 2.

For the purpose of making the variables reflect the reality directly, we define $W_i^{behav} \approx W_i^*$ as the actual (behavioural) contribution of individuals, and $W_i^{stated} \approx W_i^E$ as the WTP values stated in a survey. Then we test the condition by:

Prediction 1 (Differential response to observability) *An increase in observability s raises both stated WTP and behavioural contributions:*

$$\frac{dW_i^{behav}}{ds} > 0$$

and

$$\frac{dW_i^{stated}}{ds} > 0.$$

The wedge between behavioural and stated WTP decreases with observability if and only if WTP under optimisation is more sensitive to observability than compensating variation:

$$\frac{d\Delta}{ds} < 0 \quad \text{iff} \quad \frac{dW_i^{behav}}{ds} > \frac{dW_i^{stated}}{ds}.$$

Alignment occurs at any s^ satisfying:*

$$W_i^{behav}(s^*) = W_i^{stated}(s^*).$$

The derivatives of both WTP measures with respect to observability are shown to be positive in Appendix B. The model predicts that if behavioural contributions are more sensitive to visibility than stated valuation, the observed gap between the two should narrow as observability increases.

Empirically, this prediction can be tested by comparing how behavioural contributions and stated WTP respond to exogenous variation in observability across treatments or policy environments. For example, a policy intervention introducing public donor leaderboards, mandatory disclosure of contributions, or other visibility-enhancing mechanisms would generate variation in observability and allow comparison of behavioural contributions across different observability regimes. Ideally, one may estimate the stated and behavioural measures for the same individuals. Let W_{ik} denote the WTP measure for the individual i , where $k \in \{\text{stated}, \text{behavioural}\}$, and $Behaviour_k = 1$ for actual behavioural contribution and 0 for stated WTP. Then estimate

$$W_{ik} = \theta_0 + \theta_1 Behaviour_k + \theta_2 s_i + \theta_3 (Behaviour_k \times s_i) + \mu_i + u_{ik}.$$

Here, μ_i captures the fixed effects of individuals when both measures are observed for the same individual. Then,

$$\theta_3 = \frac{dW_i^{behav}}{ds} - \frac{dW_i^{stated}}{ds}$$

estimates the differential sensitivity of behavioural contributions relative to stated WTP. The model predicts that the wedge decreases with observability if $\theta_3 > 0$.

This test identifies whether the wedge narrows with observability, but it does not by itself prove that an alignment point s^* exists within the observed range. However, the alignment point serves more as an illustration than as a target we, or the policymakers, aim to achieve.

For the identification of θ_3 , it requires exogenous variation in observability s_i . In practice, this can be achieved through random assignment to different visibility conditions (e.g., disclosure vs anonymous contribution regimes) or through institutional changes that alter observability for some individuals but not others. If s_i is correlated with unobserved traits that affect responsiveness to observability, estimates of θ_3 may be biased. In such cases, an instrumental variable exploiting the exogeneity of exogenous variation would be required.

Further, recall that in the literature review we noted that many empirical and experimental researchers have attributed behavioural and stated WTP divergence to hypothetical bias and other constraints (Kahneman and Knetsch, 1992; List and Gallet, 2001, etc.). These remain valid. Hence, empirically, this implies that differences between the two measures may reflect both structural and behavioural components. The proposed test does not, by itself, separate structural differences from

hypothetical bias, but nor does it need to; the theoretical prediction concerns the differential response to observability, not the level difference.

To sum up this subsection, we proposed a testable prediction that links theoretical findings to an empirical implication. By running the test, we can estimate the sensitivity of WTP to observability and therefore determine whether increasing observability can reduce the value-action gap. Nevertheless, many sources of bias, such as hypothetical bias, moral expression, strategic misrepresentation, could still affect the test result. Hence, the empirical design should be interpreted as testing the mechanism of differential sensitivity, rather than providing a complete decomposition of the value-action gap.

4. Summary and Implications

This paper provides a structural interpretation of the WTP gap in climate mitigation. Rather than viewing the divergence between stated valuation and behavioural contributions solely as a consequence of bias or inconsistency, we show that the gap can arise even under rational behaviour and internally consistent preferences. The central reason is that welfare-based WTP and behavioural contribution are different economic concepts. Compensating variation WTP is defined by a level condition that equates utility across states, whereas behavioural WTP is determined by marginal conditions of an optimisation problem. Except under knife-edge conditions, there is no reason for these two concepts to coincide.

This distinction changes how the value-action gap should be interpreted. A positive gap between stated WTP and actual contribution need not imply that survey respondents are simply exaggerating their preferences, or that observed behaviour reveals a “truer” valuation. Instead, stated WTP may be a valid welfare measure, but still a poor predictor of voluntary contribution behaviour. Conversely, actual contributions may accurately reflect marginal behavioural incentives but fail to capture the full welfare value that individuals attach to climate mitigation. The empirical gap is, therefore, not necessarily a measurement failure; it may reflect a mismatch between the object elicited by the survey and the object generated by behaviour. The analysis also highlights the role of observability as a mechanism linking valuation and behaviour. When pro-environmental actions are more visible, individuals receive reputational incentives to contribute. These benefits enter the behavioural problem via marginal incentives, thereby shifting behavioural WTP. If behavioural responses are more sensitive to observability than welfare-based valuation, the gap between the two measures narrows. This model, therefore, provides a microfoundation for policies that use visibility, public disclosure, or commitment mechanisms to align voluntary behaviour with stated valuation. At the same time, excessive observability may lead to over-contribution, in which individuals contribute more than their welfare-based valuation would justify.

The framework also has implications for stated-preference survey design. If the purpose of a survey is welfare evaluation, a compensating variation interpretation of stated WTP is probably appropriate. However, if the purpose is to predict actual contribution behaviour, the survey must elicit a behavioural object rather than a purely welfare-based valuation. This distinction offers a structural interpretation of cheap-talk scripts. Cheap talk script is not modelled here as a complete theory in itself, but it can be understood as a partial object-alignment device: by reminding respondents of budget constraints and foregone consumption, it may shift responses away from a level-comparison frame and towards the marginal-choice frame that governs actual behaviour.

Finally, the paper suggests an empirical implication. The key test is not merely whether stated and behavioural WTP differ in level, since such differences may reflect both structural and behavioural forces. Instead, the model predicts that observability should affect the two concepts differently. Hence, estimating whether behavioural contributions respond more strongly to observability than stated valuations offers a direct test of the differential-sensitivity mechanism. Future work could extend the reduced-form equilibrium feedback into a fully dynamic model in which perceived environmental effectiveness evolves with past aggregate contributions and social observability.

Overall, this paper argues that the value-action gap should not automatically be interpreted as evidence that one measure is biased and the other is correct. Welfare valuation and behavioural contribution answer different economic questions. Recognising this distinction is essential for the interpretation of stated-preference data, the design of surveys, and the evaluation of policies that seek to translate climate concern into actual contribution behaviour.

5. Extensions

Despite the main results in this paper, we noted that two topics could be further developed. First, the framework developed in this paper can be extended within a “three-stage modelling” structure that progressively enriches the relationship between valuation and behaviour. The baseline model considers a static comparison between compensating-variation WTP and behavioural contribution in the absence of equilibrium feedback. In this setting, the divergence arises purely from the structural distinction between a level and a marginal condition (as presented in Sections 3.2 to 3.4).

The second stage introduces equilibrium feedback in a reduced-form manner. The valuation environment is assumed to depend on an exogenous benchmark level of aggregate contributions, capturing the hypothetical nature of survey responses. This allows the model to incorporate the idea that individuals evaluate environmental improvements as a hypothetical environment, consistent with what the real-world survey typically suggests (as presented in Section 3.5).

A natural extension is to endogenise this feedback mechanism. Rather than treating the hypothetical environment as exogenous, one can interpret it as being formed by past realised behaviour. In particular, when individuals report their willingness to pay in a survey context, they may implicitly condition their valuation on the level of environmental improvement achieved in previous periods. In this interpretation, the equilibrium benchmark introduced in the second stage corresponds to the lagged optimal behavioural contribution. We suggest that this may better represent real-world behaviours and derive clearer sign conditions, hence providing a better theoretical foundation and/or policy implications.

Another extension that comes to mind concerns the role of survey design, especially cheap-talk scripts. The present paper treats cheap talk only as an implication of the model rather than as an object of formal analysis. A more complete extension would explicitly model how survey scripts affect the respondent’s perceived decision problem. In the current framework, standard contingent-valuation questions may elicit a welfare-based object close to compensating variation, whereas a cheap-talk design that emphasises budget constraints, opportunity costs, and real-payment consequences may shift respondents towards the behavioural-choice object.

More broadly, the paper reflects a wider theoretical interest: survey-design tools such as cheap talk should not be treated only as practical corrections for empirical bias but can also be developed context by context into theoretical mechanisms for understanding stated-preference behaviour.

References

- Alpizar, F., Carlsson, F. and Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92 (5-6), 1047–1060. <https://doi.org/10.1016/j.jpubeco.2007.11.004>
- Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100, 464–477. <https://doi.org/10.2307/2234133>
- Andreoni, J. and Bernheim, B.D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5), 1607–1636. <https://doi.org/10.3982/ECTA7384>
- Ariely, D., Bracha, A. and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1), 544–555. <https://doi.org/10.1257/aer.99.1.544>
- Bartling, Engl, and Weber (2015). Game form misconceptions are not necessary for a willingness-to-pay vs willingness-to-accept gap. *Journal of the Economic Science Association*, 1(1), pages 72–85. <https://doi.org/10.1007/s40881-015-0008-0>
- Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678. <https://doi.org/10.1257/aer.96.5.1652>
- Bernheim, B.D. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1), 51–104. <https://doi.org/10.1162/qjec.2009.124.1.51>
- Carson, R.T. and Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics*, 37(1), 181–210. <https://doi.org/10.1007/s10640-007-9124-5>
- Carson, R.T. (2012). Contingent Valuation: A Practical Alternative When Prices Aren't Available, *Journal of Economic Perspectives*, 26(4), 27–42. <https://doi.org/10.1257/jep.26.4.27>
- Cummings, R.G. and Taylor, L.O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review*, 89(3), 649–665. <https://doi.org/10.1257/aer.89.3.649>
- DellaVigna, S., List, J.A. and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127(1), 1–56. <https://doi.org/10.1093/qje/qjr050>
- Hanemann, M.W. (1984). Welfare evaluations in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics*, 66(3), 332–341. <https://doi.org/10.2307/1240800>
- Hicks, J.R. (1939). The foundations of welfare economics. *The Economic Journal*, 49(196), 696–712. <https://doi.org/10.2307/2225023>
- Kahneman, D. and Knetsch, J.L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, 22, 57–70. [https://doi.org/10.1016/0095-0696\(92\)90019-S](https://doi.org/10.1016/0095-0696(92)90019-S)
- List, J.A. and Gallet, C.A. (2001). What experimental protocol influences disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, 20, 241–254. <https://doi.org/10.1023/A:1012791822804>
- Manski, C.F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*. 72(4), 1221–1246. <https://doi.org/10.1111/j.1468-0262.2004.00530.x>
- Murphy, J.J., Stevens, T. and Weatherhead, D. (2005). Is cheap talk effective at eliminating hypothetical bias in a provision point mechanism? *Environmental and Resource Economics*, 30, 327–343. <https://doi.org/10.1007/s10640-004-4224-y>
- Nordhaus, W.D. (2017). Revisiting the social cost of carbon. *Proceedings of the National Academy of Sciences*, 114(7), 1518–1523. <https://doi.org/10.1073/pnas.1609244114>
- Sen, A.K. (1977). Rational fools: A critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs*, 6, 317–344.
- Zhao, J. and Kling, C.L. (2004). Willingness to pay, compensating variation, and the cost of commitment. *Economic Inquiry*, 42, 503–517. <https://doi.org/10.1093/ei/cbh077>

Appendix A: Omitted Proofs

Proof of Proposition 1.1. Let V_i^0 denote utility in the initial state, which is purely exogenous. Then, the compensating variation WTP under equilibrium, W_i^E , is defined by

$$U(W_i^E) = V_i^0. \quad (4)$$

The behavioural optimum W_i^* is again determined by

$$U_W(W_i^*) = 0. \quad (5)$$

By strict concavity (A2), $U_W(W_i) = 0$ has a unique solution W_i^* and $U(W_i) = V_i^0$ has at most one solution W_i^E . Assume, for contradiction, that

$$W_i^* = W_i^E = \hat{W}.$$

Then, $\hat{W}$ must satisfy both equations (4) and (5):

$$\begin{aligned} U(\hat{W}) &= V_i^0 \\ U_W(\hat{W}) &= 0, \end{aligned}$$

which implies

$$U_W(W_i^E) = 0. \quad (6)$$

Equation (6) implies that the derivative is zero on a given level set, which is an additional scalar restriction on the primitives.

Assumptions A1-A2 do not imply this restriction. Hence, except for knife-edge conditions, generically

$$W_i^* \neq W_i^E.$$

Q.E.D.

Proof of Lemma 2. Under Lemma 1, the utility function $U(W_i; s)$ is strictly concave in W_i . Therefore

$$U_{WW}(W_i; s) < 0.$$

Because $U_W(W_i; s)$ is strictly decreasing:

Case 1: $w > W_i^*(s)$

Since the derivative is decreasing,

$$U_W(w; s) < U_W(W_i^*(s); s) = 0.$$

Case 2: $w < W_i^*(s)$

Similarly, by monotonicity of the derivative,

$$U_W(w; s) > U_W(W_i^*(s); s) = 0.$$

Thus

$$w > W_i^*(s) \Leftrightarrow U_W(w; s) < 0, \quad w < W_i^*(s) \Leftrightarrow U_W(w; s) > 0 \quad (7)$$

Substituting $w = W_i^E(s)$ to the two conditions in (7), we get

$$W_i^E(s) > W_i^*(s) \Leftrightarrow U_W(W_i^E(s); s) < 0, \quad W_i^E(s) < W_i^*(s) \Leftrightarrow U_W(W_i^E(s); s) > 0.$$

Since we defined the wedge as

$$\Delta(s) = W_i^E(s) - W_i^*(s),$$

$\Delta(s) > 0$ implies $U_W(W_i^E(s); s) < 0$ and $\Delta(s) < 0$ implies $U_W(W_i^E(s); s) > 0$, i.e.,

$$\Delta_i(s) \leq 0 \Leftrightarrow U_W(W_i^E(s); s) \geq 0$$

as required.

Q.E.D.

Proof of Lemma 3. Define the behavioural first-order conditions as

$$F(W, s) = U_W(W; s).$$

At an interior optimum $W^*(s)$,

$$F(W^*(s), s) = 0.$$

By A1, U is twice continuously differentiable, so F is continuously differentiable. By A2 and strict concavity in W ,

$$F_W(W^*(s), s) = U_{WW}(W^*(s); s) < 0.$$

Hence $F_W \neq 0$. By the Implicit Function Theorem, there exists a locally unique, continuously differentiable function $W^*(s)$.

Therefore $W^*(s)$ is continuous in s .

Now consider the compensating variation WTP, $W^E(s)$. Let

$$G(W, s) = u(c - W) + E(pQ(s)) + M(\bar{W}(s), s) - V^0(s),$$

where $V^0(s)$ denotes the utility in the initial state. The equilibrium-based compensating variation $W^E(s)$ is defined by

$$G_W(W, s) = -u'(c - W).$$

By A2, $u'(c - W) > 0$. Hence

$$G_W(W, s) < 0.$$

Thus $G_W \neq 0$ at $W^E(s)$. By the Implicit Function Theorem, $W^E(s)$ is locally continuously differentiable, and therefore continuous in s .

Since both $W^*(s)$ and $W^E(s)$ are continuous, the wedge

$$\Delta(s) = W^E(s) - W^*(s)$$

is also continuous in s .

Q.E.D.

Proof of Proposition 3. Let

$$\Delta(s) = W_i^E(s) - W_i^*(s).$$

By assumption, Δ is continuous and $\Delta(s^*) = 0$. Since s^* is a simple crossing point, there exists $\varepsilon > 0$ such that

$$\Delta(s) > 0 \text{ for all } s \in (s^* - \varepsilon, s^*),$$

and

$$\Delta(s) < 0 \text{ for all } s \in (s^*, s^* + \varepsilon).$$

Therefore, for s sufficiently close to s^* , we have:

- (i). $s < s^*$, then $\Delta(s) > 0$, so $W_i^*(s) < W_i^E(s)$, implying under-contribution;
- (ii). $s = s^*$, then $\Delta(s) = 0$, so $W_i^*(s) = W_i^E(s)$, implying valuation-consistency;
- (iii). $s > s^*$, then $\Delta(s) < 0$, so $W_i^*(s) > W_i^E(s)$, implying over-contribution.

Hence, the three local regimes hold around s^* .

Q.E.D.

Appendix B: Sensitivity to Observability

In this Appendix, we show that $(dW^*)/ds$ and $(dW_i^E)/ds$ exist and are nonzero.

First, consider behavioural optimisation. Differentiating the FOC $U_w(W_i^*) = 0$ with respect to s and applying the Implicit Function Theorem yields

$$\frac{dW^*}{ds} = -\frac{U_{W_s}}{U_{WW}}.$$

Assume that the observability strengthens the marginal reputation return to contributing, i.e., $U_{W_s} = M_{W_s}^{social} > 0$. One reasonable functional form (which we also used in Section 3.6) can be $M^{social}(s_i) = \lambda_i s_i W_i$, which satisfies the assumption ($U_{W_s} = \lambda_i > 0$). Then

$$\frac{dW^*}{ds} > 0.$$

Thus, behavioural WTP increases with observability.

Next, consider the CV. Differentiating W_i^E with respect to s gives

$$\frac{dW_i^E}{ds} = \frac{E'(pQ(s))p \frac{dQ}{ds} + \frac{dM}{ds} + M_w \frac{d\bar{W}}{ds}}{u'(c)}.$$

Because the equilibrium contributions $\bar{W}$ are taken as exogenous in the CV framework,

$$\frac{dQ}{ds} = 0, \quad \frac{d\bar{W}}{ds} = 0.$$

Hence

$$\frac{dW_i^E}{ds} = \frac{1}{u'(c)} \cdot \frac{dM}{ds} > 0.$$

Thus, observability increases both behavioural contributions and welfare-based valuation.

Q.E.D.