Friedrichsen, Jana; Engelmann, Dirk

Who cares about social image?

European Economic Review

This is the **postprint** of an article published in European Economic Review 110 (2018), S. 61-77
available online at: https://doi.org/10.1016/j.euroecorev.2018.08.001

Who cares about social image?

Jana Friedrichsen* Dirk Engelmann†

June 17, 2018

Abstract

This paper experimentally investigates how concerns for social approval relate to intrinsic motivations to purchase ethically. Participants state their willingness-to-pay for both a fair trade and a conventional chocolate bar in private or publicly. A standard model of social image predicts that all participants increase their fair trade premium when facing an audience. We find that the premium is indeed higher in public than in private. This effect, however, is driven by participants who preferred a conventional chocolate bar over a fair trade one in a pre-lab choice. For those who chose the fair trade chocolate bar, public exposure does not change the fair trade premium. This is captured by a generalized model where intrinsic preferences and the concern for social approval are negatively correlated.

JEL-codes: C91, D12
Keywords: image concerns; ethical consumption; fair trade; social approval; experiments

*Corresponding author. Deutsches Institut für Wirtschaftsforschung (DIW) and Humboldt-Universität zu Berlin, Mail: DIW, Mohrenstr. 58, 10117 Berlin, Germany. Email: jfriedrichsen@diw.de

†Humboldt-Universität zu Berlin; CERGE-EI, Prague; CESifo, Munich, Mail: Humboldt-Universität zu Berlin, Spandauer Str. 1, 10099 Berlin, Germany. Email: dirk.engelmann@hu-berlin.de

Accepted manuscript, published in European Economic Review 110 (2018), pp. 61-77.

1 Introduction

There is substantial evidence that individuals desire to signal pro-social preferences in the lab (e.g. Andreoni and Bernheim, 2009; Ariely et al., 2009) and in the field (e.g. Carpenter and Myers, 2010; Soetevent, 2011). Specifically, empirical research also suggests that choices for products considered to be ethical are driven not only by intrinsic motivation but also by concerns for social approval (Griskevicius et al., 2010; Sexton and Sexton, 2014). How intrinsic motivation and image concerns interact is important for the supply and pricing policy of these products (Friedrichsen, 2018). Furthermore, the interaction is relevant for assessing potentially negative effects of incentives on behavior motivated by intrinsic motivation or image concerns (Bénabou and Tirole, 2006; Gneezy and Rustichini, 2000).[1] Little is known about this interaction between intrinsic motivation and concerns for social approval from empirical work. This paper starts filling this gap by assessing in a laboratory experiment, how intrinsic motivation and social image concerns interact.

In the experiment, we test whether individuals who are intrinsically motivated to support fair trade react more or less strongly to opportunities for image building than individuals who are not intrinsically motivated. Before subjects come to the experiment, we elicit a proxy for their intrinsic motivation for supporting fair trade by offering a choice between two chocolate bars: either a smaller fair trade one or a larger conventional one. This is offered as an additional reward for taking part in the experiment. In the experiment itself, participants first engage in a market game unrelated to chocolate or fair trade products. Then we elicit in an incentive-compatible way their willingness-to-pay both for conventional and fair trade chocolate in order to derive the fair trade premium each individual subject is willing to pay, which is our main variable of interest. Our treatments vary whether the willingness-to-pay is kept private or must be announced publicly. Thus, we vary whether participants can build an image among their fellow participants for being concerned about fair trade. In a questionnaire, we ask

[1] Consider a supposedly ethical product (such as environmentally friendly, fair trade, or those bundled with charitable donations) that is purchased by people who are intrinsically motivated to support the cause related to the product as well as those primarily concerned with the positive image derived from this purchase. Providing extrinsic incentives can dilute the image derived from the purchase and, as a result, crowd out purchases of image-motivated consumers. Crowding out of intrinsically motivated consumers may be problematic because they are arguably more willing to make an effort to observe whether the products satisfy the necessary standards. Intrinsically motivated agents are actually crowded out only if they are also concerned with their social image.

for knowledge about and attitudes toward fair trade and confirm that the latter are positive, so that our variation of image building opportunities indeed allows participants to derive a positive social image.

Our theoretical framework follows Bénabou and Tirole (2006) in assuming that behaving pro-socially confers a positive social image or social esteem, such that image-concerned individuals should behave more pro-socially in the public sphere than in private.[2] Under the standard assumption that all individuals have identical image concerns, we derive several hypotheses that we test in our experiment. Our main hypotheses are, first, that in the private condition, participants who chose a fair trade chocolate bar before the experiment will state an average fair trade premium that is higher than the one stated by those participants who chose a conventional bar before the experiment (Hypothesis 1), second, that all participants will increase their fair trade premium by the same amount if it is elicited publicly than if it is elicited in private (Hypothesis 2) and third, that as a consequence also in the public condition those who chose fair trade chocolate before the experiment will state a higher fair trade premium than those who chose conventional chocolate (Hypothesis 3).

In the private setting, we find that subjects who revealed that they have no intrinsic motivation for fair trade by choosing the conventional chocolate bar before the experiment began have significantly lower fair trade premiums than those who revealed an intrinsic fair trade preference by choosing the fair trade chocolate bar, supporting Hypothesis 1. In contrast to Hypothesis 2, though, only those participants who chose the conventional chocolate before the experiment exhibited a significantly larger fair trade premium in public than in private. As a result, in the public setting on average both groups of participants stated fair trade premiums that were not statistically different, thus not supporting Hypothesis 3.

Using a generalized model of image concerns, we show that all of our experimental findings are rationalized if intrinsic motivation and image concerns are negatively correlated and present a possible foundation for such a negative correlation. Furthermore, we discuss alternative approaches and show that our data is inconsistent with models of fair trade purchases that predict a positive correlation, as, for instance, a model of expressive behavior does.

[2]Alternative signaling explanations for charitable behavior argue that the size of the contribution reveals an individual's wealth (Glazer and Konrad, 1996; Harbaugh, 1998). This channel is absent from our design since purchases were made from the experimental endowment, but it may play a role in the field evidence on image concerns in giving.

These findings have three implications. First, incentives (such as, for example material rewards) that encourage those not intrinsically motivated to engage in socially beneficial behavior–which, as argued above, can undermine the image from the corresponding choice–will not crowd out intrinsically motivated socially beneficial behavior in our sample. This finding suggests that incentivizing fair trade purchases (e.g., through subsidies) may be a reasonable idea. Second, as a general insight, correlations between intrinsic motivation and image concerns are economically relevant, and theoretical models therefore need to allow for significant correlations between types and their signaling concerns. Our stylized extended model suggests that this is feasible in a tractable way. Third, for the inferred negative correlation between intrinsic motivation and image concerns, an optimally designed product portfolio will induce partial pooling of different consumer types (cf. Friedrichsen, 2018).

Our experiment uses fair trade products, understood as bundles of a base product with a charitable contribution as in Reinstein and Song (2012).[3] By stating a high willingness-to-pay for the bar of fair trade chocolate in the experiment, the participants can signal their pro-social attitudes. This is natural because consumer products are a frequently used signaling medium (e.g. Miller, 2009). Teyssier et al. (2015) find that the willingness-to-pay for conventional chocolate decreases in public, leading to higher premiums for fair trade in public, but they did not investigate intrinsic preferences or heterogeneity. In addition, a field experiment in German coffee shops finds that consumers donate larger amounts if these are bundled with a coffee purchase than if coffee purchases and donations are separated (Koppel and Schulze, 2013). This is explained by social image concerns because the consumers have to communicate their choices to the staff if the donations are bundled with the product, whereas direct donations are just dropped into a box anonymously. As this paper focuses on consumer preferences and behavior, we do not discuss potential benefits or pitfalls of fair trade but refer to Dragusanu et al. (2014) and the literature surveyed therein for an economic analysis of it.[4]

We introduce our design in Section 2, develop a theoretical model to derive our hypotheses in Section 3, and present our results in Section 4. In Sections 5 and 6, we

[3]Such bundles need not be efficient, i.e., the price of the bundle may exceed the sum of the price of the private good and the donation as would be the case if retailers increase their markups for fair trade products. Consumers indeed also choose inefficient bundles (see Munro and Valente, 2016, for experimental evidence).

[4]A literature survey focusing on the effects of fair trade on the welfare of smallholders is provided by Dammert and Mohan (2015).

discuss possible explanations and implications. Section 7 concludes. Appendices A, B, and C provide proofs and robustness results for our model and are contained at the end of the paper. Appendices D, E, and F are available as an online supplement and contain additional regression results and tests as well as the instructions.

2 Experimental design and procedures

Our experimental design consists of two sets of choices. First, after participants have registered for the experiment, but before they arrive at the laboratory, we derive a proxy to determine their preference for fair trade. Second, they participate in a laboratory experiment in which they take part in a market game before we elicit their willingness to pay a price premium for a fair trade product.

Proxy for intrinsic preference In order to derive a proxy for their intrinsic preference for fair trade, via email we offered subjects a choice between a bar of fair trade milk chocolate and a bar of conventional milk chocolate as an additional reward for coming to the experiment. This email was sent and had to be answered before they came to the laboratory, but the chocolate bars were distributed only after the experiment. Since fair trade chocolate is typically more expensive, we offered a choice between a slightly larger (125g) bar of conventional chocolate and a standard size (100g) bar of fair trade chocolate.[5] As few subjects chose the conventional chocolate bar in our first sessions, we offered a choice between two bars of conventional chocolate and one bar of fair trade chocolate in the following sessions. We balanced the design with respect to whether we offered one or two bars of conventional chocolate. The difference in trade-offs offered should have moved some subjects with a positive willingness-to-pay to support fair trade between categories. This can only have weakened the comparison between the two classes of subjects.[6] Another reason why our proxy could be noisy is that some participants might not like milk chocolate at all. We expect that even in this case they would choose the chocolate produced according to their preferred method, because even if they give away or throw away the chocolate it makes sense to choose the fair trade bar if they think fair trade is helpful and choose the conventional one if they think fair trade is a scam. They might also, however, simply randomize because

[5] An English translation of the email text can be found in Appendix E.

[6] The main results hold for either version of the trade-off. See also footnotes 22 and 23.

they do not think much about the production standard of a product they would never buy. In this case, on expectation half of them would be classified wrongly and thus our classification would be diluted. This could weaken the differences between the fair trade and conventional choosers as we classify them.[7]

Market game The laboratory experiment itself consists of two parts. In the first part, the participants take part in a market game modified from Danz et al. (2012), with participants taking the roles of firms, consumers and workers. This market game serves two purposes. On the one hand, given that the second part is short, we used the opportunity to assess the generalizability of fair behavior observed in an experimental market by comparing behavior in this market with fair trade choices both before the experiment and in its second part. This analysis is the focus of a separate paper (Engelmann et al., 2018).[8] On the other hand, the market game serves to start the experiment in a relatively conventional fashion, thus removing the focus from the rather unusual chocolate purchase in the second part. This is particularly important as the participants were promised a bar of chocolate as an additional reward for participating in the experiment. If the experiment had started with another chocolate-related choice, subjects might have connected the two chocolate-related choices to each other. We think such a connection that might trigger undesired demand effects is less likely with the chosen order. Indeed, many subjects had forgotten about the first (email-based) chocolate when they were paid at the end of the experiment and we intentionally did not remind them during the experiment that they already had made a chocolate choice before the experiment.

Willingness-to-pay In the second part, we use a random price mechanism (Becker et al., 1964) to elicit from each participant his or her willingness-to-pay (WTP) for both a fair trade and a conventional dark chocolate bar. Specifically, subjects enter a price

[7]Furthermore, participants who have a very weak preference in favor of fair trade that would normally not outweigh the size difference between the two types of chocolate that we offer might choose the fair trade bar if they intend to give it away and hence be classified as fair trade choosers but would not if we had offered a taste variant that they liked. Again, this should only dilute the difference between the two groups of participants.

[8]In this market game, a worker player passively collects wages, two firms choose prices and wages and a monopsonistic buyer who is informed about both prices and wages can decide how to split purchases between the two firms. The passiveness of the worker player combined with the market power of the buyer make this setting suitable to study the buyer's willingness to pay for fair treatment of workers in a market setting.

between 0 and 2 Euros, where any multiple of €0.01 is permitted. Then we draw a price from a uniform distribution of all integer multiples of €0.01 between 0 and 2 Euros. Subjects receive a bar of the chocolate type sold if their stated WTP for that type is at least as high as the randomly chosen price. Which type of chocolate is sold is determined randomly after the price has been chosen such that the mechanism is incentive compatible for both types of chocolate. We chose dark chocolate for this part of the experiment instead of milk chocolate. This ensures that subjects cannot end up with two identical bars of chocolate, which could have reduced their willingness-to-pay for the type of chocolate that they were already sure to receive. We also did not choose well-known brands, in order to minimize the chance that the subjects' willingness-to-pay was based on taste preferences due to personal experience or anchoring on market prices. From these two WTPs, we infer individuals' willingness' to pay a premium for the fair trade chocolate as $\text{WTP}_{\text{fair}} - \text{WTP}_{\text{conv}}$. We call this the *fair trade premium*.

Treatments Our two treatments differ in whether the WTPs are elicited publicly or in private. In treatment *private*, individuals enter their WTPs privately at the computer. In treatment *public*, after they have entered their WTPs privately at the computer, all subjects stand up and announce their WTPs publicly among the group of participants. They are informed about this procedure before entering their WTPs at the computer. The difference in the fair trade premiums between the treatments serves as our measure for image concerns.

We note that while the random price mechanism (Becker et al., 1964) is incentive compatible in theory, it has been pointed out that experimental subjects may misconceive this mechanism (Plott and Zeiler, 2005; Cason and Plott, 2014). Such misconceptions should be of much less concern in our experiment. Misconceptions appear to be more of an issue for elicitation of willingness to accept to forego an item rather than for willingness to pay to obtain an item. More importantly, we are only interested in the fair trade premium and, in particular, in the question of whether this differs significantly across groups or treatments. Our analysis is robust to misconceptions that only lead to a bias that is monotone in the true WTP.[9] One could imagine further misconceptions, but it appears that in order to test hypotheses relying on whether differences (as well as differences-in-differences and diffs-in-diffs-in-diffs) in WTP are significant, the

[9]Note that monotonicity implicitly requires that the misconception is not systematically correlated with the treatment or the chocolate choice before the experiment.

random price mechanism is substantially more robust to misconceptions than when it is used to measure absolute WTP.

Procedures The experiment was computerized using zTree (Fischbacher, 2007) and took place in the experimental economics laboratory at the University of Mannheim (mLab) in May, June, and October, 2012. Participants were recruited using ORSEE (Greiner, 2015). An English translation of the (German) instructions for the second part of the experiment is included in Appendix F. We conducted 8 sessions with 16-20 participants each, with a total of 144 participants. For the market game, each participant received a show-up fee of €5; for the second part of the experiment in which we elicit the WTPs, each participant received an additional endowment of €4. Average cash earnings were €18.63, including the show-up fee and the endowment in the second part, as well as subtracting payments for chocolate if applicable.[10] In the second part, the payoff-relevant chocolate was conventional and fair trade in half of the sessions each. We handed out conventional chocolate to 22 subjects and fair trade chocolate to 21 subjects. Details about the (randomly chosen) prices at which chocolates were sold are collected in Table 1.

After entering their WTPs (but before they announce them in *public*), subjects fill in an extensive questionnaire regarding their attitudes toward and knowledge about fair trade. The answers to this questionnaire allow us to confirm the validity of our proxy for intrinsic motivation as those who chose fair trade chocolate report buying fair trade products significantly more often and report significantly more positive attitudes toward fair trade (see Figures 5 and 6, and Table 3 in Appendix D.1).

Participating in the market game could potentially impact the stated willingness-to-pay in the last part of the experiment. Note, however, that up to and including the market experiment, the two treatments are perfectly identical and hence there is no clear reason how the market game should lead to treatment differences and moreover to heterogeneity in the reaction to the treatment. One plausible route how the market game could have impacted an individual's willingness-to-pay is through the earnings from this stage. We thus control for these earnings in the regression analysis below.

[10] In the market game in the first part of the experiment, participants in the role of firms earned €4.50 on average, those in the role of workers earned €6.31 on average, and those in the role of consumers earned €23.73 on average.

	conventional				fair trade			
price in €	0.26	0.27	0.97	1.85	0.25	1.01	1.20	1.78
treatment	public	private	private	public	public	private	public	private
#participants	16	16	20	20	16	16	20	20
#bars sold	11	9	2	0	12	2	5	0

Table 1: Prices drawn and number of chocolate bars paid out to participants.

Our analysis evaluates the decisions of 121 subjects who made their choice between fair trade and conventional chocolate via email as described above. Among these 121 subjects, before coming to the lab, 32 chose conventional chocolate, while the remaining 89 chose fair trade. Both types of chocolate choosers are distributed almost evenly across treatments: in the private treatment, we have 14 subjects who chose the conventional and 46 who chose the fair trade chocolate; in the public treatment, we have 18 subjects who chose the conventional and 43 who chose the fair trade chocolate. In addition, 23 newly recruited subjects participated in our experimental sessions, but are not included in the analysis. For these subjects, the chocolate choice that we intended to use as a proxy for their intrinsic preference was taken in public during a recruitment day and not via email.[11]

3 Theoretical background and hypotheses

In this section, we derive hypotheses based on a standard model of image concerns where all consumers have identical concerns for social image. We then consider the implications of heterogeneous image concerns. Proofs are in Appendix A.

Preference Suppose utility of consuming conventional and fair trade chocolate are given by

$$v_{\text{conv}}(p_{\text{conv}}) = uq_{\text{conv}} - p_{\text{conv}} \text{ and } v_{\text{fair}}(p_{\text{fair}}) = uq_{\text{fair}} + m - p_{\text{fair}} \tag{1}$$

[11]Out of 222 new recruits, only 23 actually showed up to one of our experimental sessions and their chocolate choices are not balanced across treatments such that we cannot separately analyze this subgroup. A pooled analysis is not sensible because of the public recruitment situation, which biases the proxy due to image concerns already being active at this stage.

respectively, where p_{conv} and p_{fair} denote the prices of conventional and fair trade chocolate, u is the marginal utility of chocolate, q_{conv} and q_{fair} are the (potentially quality-adjusted) sizes of the conventional and the fair trade chocolate bar, and m is an individual's intrinsic motivation for supporting fair trade.

Assume that u is normally distributed with mean $\bar{u}$ and variance σ_u. Further, assume that preferences for fair trade, m, are normally distributed with mean $\bar{m}$ and variance σ_m, and that these are independent of the taste for chocolate u. Each individual privately knows her realizations of u and m.

Classification by proxy Denote the size of the conventional chocolate by $q_{\text{conv}} = x$ and the size of the fair trade chocolate by $q_{\text{fair}} = x - d$, where $d > 0$. Individuals select the fair trade chocolate if

$$(x - d)u + m > xu \tag{2}$$

and choose conventional chocolate if the reverse inequality holds.[12] An individual is indifferent between the fair trade bar and the larger conventional chocolate bar if her intrinsic motivation to support fair trade exactly compensates for the reduction in quantity. Thus, the threshold valuation of fair trade that compensates the size difference at a given chocolate valuation u is defined as an increasing function $m(u) = du$. Given the independent distribution of m and u, those classified as fair trade choosers are expected to have a higher m and a lower u on average. We find, however, that the fair trade and conventional choosers do not differ in u, see Appendix B.[13]

Willingness-to-pay in the private setting If asked for her WTP, an individual may deviate from (1) by a random noise term. Moreover, for clarity of exposition, we abstract from the effect of the size of the chocolate bar on WTP and normalize the size to 1.[14] In the public setting, stated WTP can also be affected by image concerns. We will introduce these below after discussing WTP in the private setting. Specifically, we assume that the stated WTPs for both types of chocolate are given by

[12] We note that in the classification the prices are irrelevant because in this stage, participants can choose one type of chocolate without having to pay.

[13] One reason for this observation may be that fair trade chocolate is typically perceived to be of higher quality, which is also supported by our questionnaire data. Thus, the perceived quality difference may compensate for the size difference.

[14] In principle, the size difference may also affect stated WTP, and the model can be extended to cover this without affecting our hypotheses regarding the treatment. See Appendix B for evidence that we do not find an effect of the size difference on the WTP.

$$w_{\text{conv}}(u, m) = u + \eta_{\text{conv}} \text{ and } w_{\text{fair}}(u, m) = u + m + \eta_{\text{fair}}, \tag{3}$$

where $\eta_{\text{conv}} \sim \mathcal{N}(0, \sigma_{\text{conv}})$ and $\eta_{\text{fair}} \sim \mathcal{N}(0, \sigma_{\text{fair}})$.[15] Since heterogeneous tastes for chocolate u have a level effect on both WTPs, we focus on the individuals' willingness to pay a premium for fair trade chocolate as given by

$$a(m) := w_{\text{fair}}(u, m) - w_{\text{conv}}(u, m) = m + \varepsilon \tag{4}$$

where $\varepsilon = \eta_{\text{fair}} - \eta_{\text{conv}}$ is a noise term that is normally distributed with $\varepsilon \sim \mathcal{N}(0, \sigma_R)$, and σ_R is the variance of the difference of both individual error terms.

As there is no scope for social signaling in the private case, only the individual's intrinsic valuation for fair trade and her utility from chocolate affect her WTP. The stated premium for fair trade is directly linked to the individual's valuation for fair trade: if $m' > m$, then $E[a|m'] = m' > m = E[a|m]$. Furthermore, upon knowing which type of chocolate an individual chose in the classification stage, we can update our expectation of her motivation to support fair trade because $E[m|\text{chose fair}] > E[m|\text{chose conv}]$. This yields the following hypothesis.

Hypothesis 1. *(Private elicitation) If elicited in private, the average fair trade premium is higher for individuals classified as fair trade choosers than for those classified as conventional choosers,* $E[a|\textit{chose fair}] > E[a|\textit{chose conv}]$.

Hypothesis 1 follows if, conditional on their basic utility from chocolate, the WTP for fair trade is higher for fair trade choosers because they have the stronger fair trade preference m, which made them choose the fair trade chocolate bar over the conventional one. Alternatively, Hypothesis 1 would also be implied if the fair trade choosers simply care less about chocolate, i.e., have a smaller u rather than a higher m. As already noted above, we do not find evidence that u differs between fair trade choosers and conventional choosers, see Appendix B.

Willingness-to-pay in the public setting Individuals may care about their social image as a pro-social person. In the private setting, there is a direct link between an in-

[15] We note that while we allow for negative valuation of chocolate ($u < 0$), negative prices are excluded in the experimental design and stated willingness-to-pay is restricted to non-negative values. An individual can always state a WTP of zero in which case receiving a chocolate bar is almost a zero-probability event. Also, chocolate can be freely discarded, which makes eliciting negative WTP non-incentive compatible.

dividual's pro-sociality as given by her preference for fair trade, m, and her fair trade premium a. This suggests that the fair trade premium can be used to signal pro-sociality.

Assume that the **social image** function $R(a) : \mathbb{R} \to \mathbb{R}$ attached to an observed fair trade premium a is the public's belief about an individual's attitude m toward fair trade. This inference function maps each observed fair trade premium, a, to an expected fair trade preference, which in equilibrium must be consistent with the conditional expectation of an individual's type upon observing her stated premium a, i.e., $R(a) = E[m|a]$ in equilibrium. Note that we need the willingness-to-pay for both types of chocolate to compute the social image. When only the willingness-to-pay for fair trade chocolate is observed, the inference on m is confounded by heterogeneity in the base utility from chocolate u. Under the assumption that w_{conv} is a reliable measure of u also in the public setting,[16] considering the fair trade premium a controls for the base utility from chocolate u.

Stated WTP thus determines an experimental participant's outcome in two ways. First, a higher WTP for a type of chocolate increases the probability that the participant will purchase this type of chocolate. Second, the stated fair trade premium a affects the participant's social image $R(a)$. Therefore the expected utility from stating the WTP pair $(w_{\text{fair}}, w_{\text{conv}})$ is

$$\begin{aligned} EU(w_{\text{fair}}, w_{\text{conv}}) &= \frac{1}{2}\left[\int_0^{w_{\text{conv}}} v_{\text{conv}}(p_{\text{conv}})dF_{\text{conv}}(p_{\text{conv}}) + \int_0^{w_{\text{fair}}} v_{\text{fair}}(p_{\text{fair}})dF_{\text{fair}}(p_{\text{fair}})\right] \\ &+ \lambda\rho R(a) \end{aligned} \quad (5)$$

where $a = w_{\text{fair}} - w_{\text{conv}}$ is the stated fair trade premium, $F_{\text{conv}}(p_{\text{conv}})$ and $F_{\text{fair}}(p_{\text{fair}})$ are the cumulative distribution functions of p_{conv} and p_{fair}, respectively and consumption utility $v_{\text{conv}}(p_{\text{conv}})$ and $v_{\text{fair}}(p_{\text{fair}})$ are given by (1). The realized social image is denoted by $R(a)$, ρ is the marginal utility of social image, and $\lambda \in \{0, 1\}$ indicates whether the purchasing decision is private (0) or public (1).[17]

Since both preferences and the error terms in the stated WTP are normally distributed, all WTP differences on the real line can occur with positive probability, and

[16] This assumption is supported by our data. Figure 2 shows that w_{conv} is virtually unaffected by the treatment.

[17] We introduce λ here so that the same utility function applies to both treatments. In principle, λ can represent a continuous measure of the degree of observability, but given that we only have two treatments, we can normalize $\lambda = 0$ in *private* and $\lambda = 1$ in *public*.

in particular all differences that can technically be observed in our experiment, i.e., all $a \in [-2, 2]$ are expected to occur with positive probability for individuals in an equilibrium. Then, the inference function, and so the anticipated social image, is well-defined across the entire interval. In line with the literature (e.g., Bénabou and Tirole, 2006), we concentrate on the case where the image function is differentiable.

Assumption 1. *Assume that the inference function $R(a)$ is differentiable in a.*

To obtain a higher observed premium and improve her social image, an individual may increase her WTP for fair trade. Stating a WTP for fair trade is costly though. Specifically, we assume that an increase in the WTP for fair trade and, thereby, the fair trade premium from the privately optimal level $a = m$ to $a = m + \Delta$, is associated with a cost $c(\Delta) = k\frac{\Delta^2}{2}$ for some scalar $k > 0$.[18] We expect individuals to increase their stated price premium in response to social image concerns as long as the marginal expected benefit exceeds the marginal expected costs, $c'(\Delta) = k\Delta$. Thus, in the public context, we expect an individual with preferences (m, u) to state a fair trade premium $a = m + \Delta + \varepsilon$ such that

$$\rho \frac{\partial R(m + \Delta)}{\partial \Delta} = k\Delta \tag{6}$$

Knowing this, an observer expects that this agent has a fair trade preference

$$E[m|a] = a - \frac{\rho}{k} \frac{\partial R(m + \Delta)}{\partial \Delta} \tag{7}$$

Solving the resulting first-order differential equation, we obtain the following result.

Proposition 1. *Let all agents have the same image concern ρ. There is a unique (differentiable-image) equilibrium, in which an agent with preferences (m, u) states a fair trade premium $a = m + \rho/k$. The marginal social image return equals ρ/k, and it is constant across types.*

Individuals of all preference types inflate their fair trade premium by increasing their WTP for fair trade by the same amount (see Figure 1). They do so in an attempt to obtain

[18] The random price mechanism in our experiment induces a quadratic cost function that is scaled by the upper bound of the price domain. When an individual increases her WTP for fair trade chocolate by Δ, then both the probability that she will buy as well as the expected price conditional on buying increase linearly in Δ and hence the expected costs are quadratic in Δ. Specifically, she faces the following additional net cost for an increase by Δ above the privately optimal level, which we here denote by w: $C(\Delta) = \int_w^{w+\Delta} \frac{p_{\text{fair}}}{\bar{p}} dF_{\text{fair}}(p_{\text{fair}}) - \frac{\Delta}{\bar{p}} w = \frac{1}{2} \frac{\Delta^2}{\bar{p}}$, where $\bar{p}$ is the upper bound on the price interval.

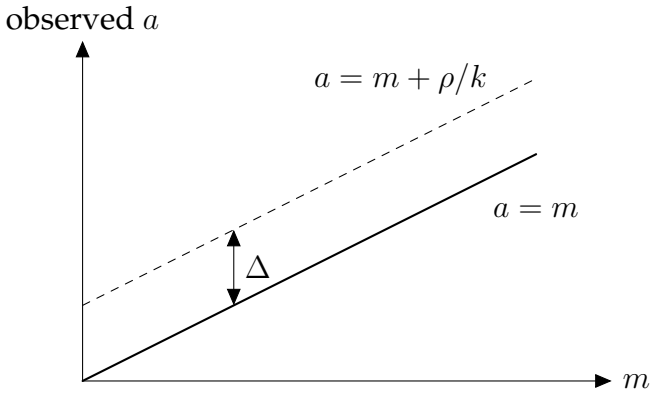

Figure 1: Expected upward shift in a due to public exposure is homogeneous (parallel) for equilibrium or naive inference according to Propositions 1 and 2.

a better social image. In equilibrium, however, everyone is just "running to keep in the same place" (Hopkins and Kornienko, 2004), and despite the increase in her fair trade premium, each type's preference for fair trade is correctly inferred. Alternatively, if individuals are unable to anticipate the equilibrium inferences but make naive inferences instead, the same incentives to inflate one's fair trade premium are at work. However, in this case everyone realizes utility from an image that exceeds her true type because of the incorrect, naive inference.

Proposition 2. *If the social image of a fair trade premium a is naively inferred as $R(a) = a$, an agent with preferences (m, u) states a fair trade premium $a = m + \rho/k$.*

The two previous results imply, that public exposure increases stated WTP of both the individuals who are intrinsically motivated and those who are not in the two benchmark cases of equilibrium inference and naive inference so that our treatment should be effective.

Hypothesis 2. *(Difference public vs. private) The average stated fair trade premium is higher if elicited publicly than privately both for conventional choosers and for fair trade choosers. In particular, the treatment difference is the same in both groups.*

As all individuals are expected to increase their WTP by the same amount, the stated fair trade premiums of conventional choosers and fair trade choosers should differ not only in the private but also in the public treatment.

Hypothesis 3. *(Public elicitation) If elicited in public, the average stated fair trade premium is higher for fair trade choosers than for conventional choosers.*

The previous propositions assume that individuals value image to the same extent. We further note that given our quadratic cost function, (6) shows that as long as the inference function is increasing, anyone with a positive concern for image would inflate her fair trade premium in the public treatment. But if individuals systematically differ in how much they value their social image, the reaction to public exposure will be heterogeneous.

Proposition 3. *If conventional choosers have systematically lower (higher) concern for social image than fair trade choosers, they increase their fair trade premium by a smaller (larger) amount in response to public exposure than the fair trade choosers.*

The intuition is the following: By (6) an individual who cares more about social image, i.e., has a larger ρ, would choose to inflate her fair-trade premium even more. Therefore, if those who are truly fair-minded cared more about their social image as suggested, e.g., by theories of expressive preferences (see, e.g., Hillman, 2010), we should see a larger treatment effect for those who are more concerned with fair trade.

Hypothesis 4. *(Expressive behavior) The treatment effect from public exposure is larger for fair trade choosers than for conventional choosers.*

4 Experimental results

In line with the results of previous studies (e.g. De Pelsmacker et al., 2005; Loureiro and Lotade, 2005), we find significant heterogeneity in the intrinsic preferences for fair trade. In our sample, the minimum fair trade premium amounted to -0.49 euros whereas the maximum fair trade premium amounted to 1.79 euros. Figure 2 illustrates the average fair trade premiums and the willingness-to-pay for both types of chocolate, separately for each group of participants and each treatment. In the following, we investigate to what extent the variation in fair trade premiums can be organized within our theoretical framework.

We first analyze whether the classification based on the chocolate choice before the experiment indeed translates into higher fair trade premiums for fair trade choosers than for conventional choosers in the private treatment. In line with Hypothesis 1, for the conventional choosers we observe an average fair trade premium $a =$ € -0.06, while for the fair trade choosers, we observe an average fair trade premium $a =$ € 0.26. The

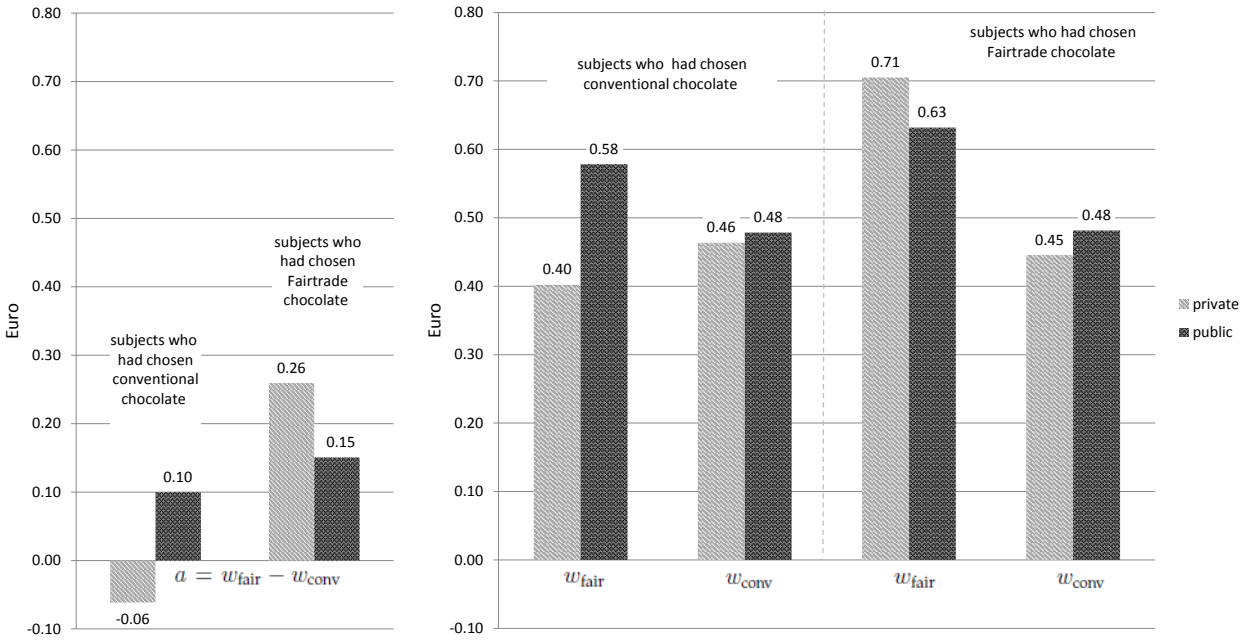

Figure 2: Averages of stated WTP by treatment and by chocolate choice. Left panel: averages of fair trade premiums a, right panel: averages of WTPs for the conventional chocolate bar, w_{conv}, and of the WTPs the fair trade chocolate bar, w_{fair}.

fair trade premiums differ across the two groups according to a Mann-Whitney test ($p < 0.001$).

Concerning the predicted treatment effect on the fair trade premium, we see a clear treatment effect in line with Hypothesis 2 on the conventional choosers. Making choices public increases the average stated fair trade premium of the conventional choosers from €-0.06 in private to €0.10 in public. The two distributions of fair trade premiums are significantly different (Mann-Whitney test, $p = 0.005$). The treatment difference is driven by an increase in the willingness-to-pay for fair trade chocolate rather than a decrease in the willingness-to-pay for conventional chocolate, w_{fair} increases from €0.40 to €0.58, but w_{conv} even marginally increases (€0.46 in private and €0.48 in public).[19]

In contrast, there is no significant treatment effect on the fair trade premium of fair trade choosers. The average stated fair trade premium of a fair trade chooser even decreases from €0.26 in private to €0.15 in public, but the decrease is not statistically

[19]The fair trade premium could, in principle, also be inflated by decreasing the WTP for conventional chocolate. This does not appear to happen, possibly because individuals bracket narrowly (Read et al., 1999) and consider the image of buying fair trade only when stating WTP for fair trade chocolate.

significant (Mann-Whitney test, $p = 0.122$).[20] Hence Hypothesis 2 is not supported for the fair trade choosers. This result is inconsistent with the assumption of identical image concerns of all participants and instead suggests that the conventional choosers are concerned with their image, while fair trade choosers are not.

As a result of the heterogeneous treatment effect, the difference between the fair trade premiums between the conventional choosers and the fair trade choosers nearly disappears in the public treatment and is no longer significant (Mann-Whitney test, $p = 0.123$). Thus, while behavior in the private treatment differs in line with intrinsic motivation, the two groups become indistinguishable in the public treatment. In particular, the fair trade premium of fair trade choosers is not higher than that of conventional choosers in public, and Hypothesis 3 is not supported. Moreover, our results contradict Hypothesis 4 (expressive behavior) because we find that individuals who chose conventional chocolate react more strongly to the treatment than the fair trade choosers. Instead, using Proposition 3, our results would be consistent with a negative correlation between intrinsic preferences for fair trade and image concerns.

On the aggregate level, for the fair trade choosers we observe a decrease in the fair trade premium in the public treatment (see Figure 2). While this is not statistically significant, such a negative treatment effect, if it was robust, could have an interesting interpretation. It would be in line with fair trade choosers choosing fair trade to support their self-image, but that the expected pooling of those only driven by social image in the public treatment leads to a decrease of the self-image derived from a given stated fair trade premium. This would result in fair trade choosers actually choosing a smaller fair trade premium in the public treatment if self-image is derived as if in the eye of a neutral observer as in the model by Benabou and Tirole (2003).[21] Such motivations do not seem to play a role here because the treatment effect on the fair trade choosers completely disappears once we control for the earnings from the first part of the experiment as will be discussed below.

The change in the distribution of fair trade premiums for the two different groups is illustrated by the cumulative distribution functions in Figure 3. On the left, we see

[20] As for the conventional choosers, the change in the fair trade premium of fair trade choosers mostly comes from a change in the WTP for the fair trade chocolate bar. For the fair trade choosers w_{fair} decreases from €0.71 in private to €0.63 in public and is therefore more strongly affected than w_{conv}, which marginally increases from €0.45 in private to €0.48 in public, see Figure 2.

[21] In Appendix C, we provide an extended model that would be consistent with a negative treatment effect for fair trade choosers.

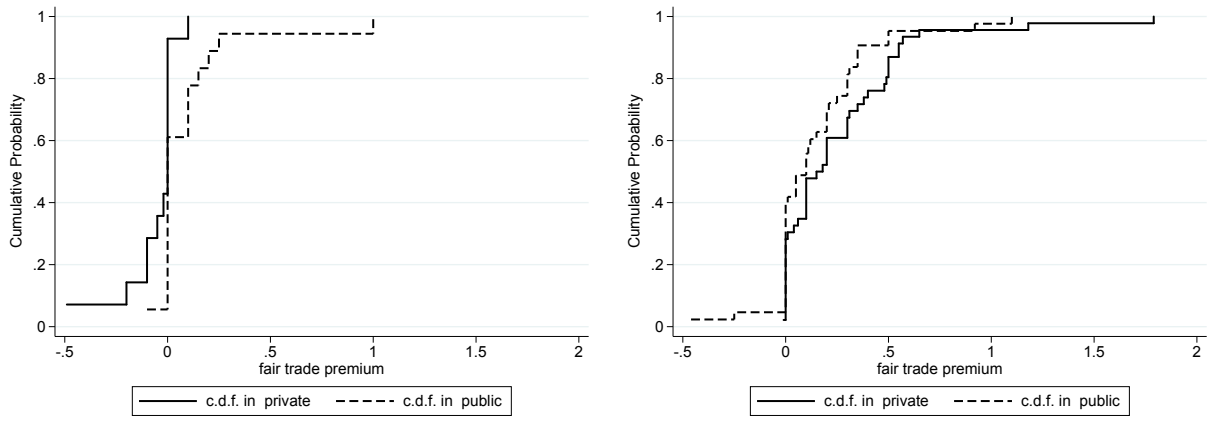

Figure 3: Cumulative distribution functions of fair trade premiums a. Left panel: conventional choosers, right panel: fair trade choosers.

that for the conventional choosers the distribution of fair trade premiums in the public treatment weakly first-order stochastically dominates that in the private treatment. Specifically, substantial mass is shifted from the left of 0 to the right of 0 when moving from the private to the public treatment. In contrast, for the fair trade choosers, fair trade premiums move in the other direction. In this case, the distribution in the private treatment nearly first-order stochastically dominates that in the public treatment. However, almost all of the mass is at or above zero in both treatments and the shift only occurs among positive premiums. As argued above, the shift in the fair trade premiums of fair trade choosers is most likely driven by differential earnings from the market game for which we control in the following regression analysis.

We confirm the heterogeneity of the treatment effect with OLS regressions of the fair trade premium on dummies for the *public* treatment and whether the subject had chosen fair trade chocolate (*FTchoice*) before coming to the experiment as well as the subject's earnings from the market game in the first part of the experiment (*marketprofit*). We also include the interaction effects between the two dummies and between earnings and the treatment dummy.

Looking at all 121 individuals, we find the following results (see Table 2, column 1). Our treatment dummy *public* is significant ($+29.2$ Cents, $p = 0.018$), thus implying that making choices public increases individuals' willingness to pay a premium for fair trade chocolate. Moreover, the heterogeneity of the treatment effect is confirmed by the negative interaction effect between having chosen fair trade chocolate and *public* ($p = 0.016$).

Table 2: Regression of the stated fair trade premium $a = w_{\text{fair}} - w_{\text{conv}}$ on (*ex-ante*) fair trade choice, treatment (public), earnings from the first part of the experiment (market-profit), and interaction terms.

	all	pos. demand
FTchoice	0.313***	0.341***
	(0.085)	(0.093)
public	0.292**	0.306**
	(0.122)	(0.133)
FTchoice*public	-0.290**	-0.289**
	(0.119)	(0.131)
marketprofit	0.007*	0.006
	(0.004)	(0.004)
marketprofit*public	-0.012**	-0.012*
	(0.006)	(0.007)
constant	-0.129	-0.122
	(0.082)	(0.090)
observations	121	104
adjusted R^2	0.113	0.113
F	4.070	3.631

Notes: Column 1 includes all subjects. In column 2, we restrict the sample to the subjects with positive demand, i.e. we exclude subjects who bid less than 2 cents for each type of chocolate. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The size of the coefficient on the interaction is similar in size to the aggregate treatment effect. Thus, in our experiment, a treatment effect indicative of image concerns is significantly less pronounced and virtually absent for those individuals who *ex-ante*, over email, chose the fair trade chocolate, in contrast to Hypothesis 2.[22] In an OLS regression of the fair trade premium restricted to the subjects who chose fair trade before the experiment, the coefficient on the dummy for the public treatment is actually positive,

[22]This result remains qualitatively unchanged if we run the analysis separately for individuals facing the choice between one fair trade chocolate bar and one, respectively two, bar(s) of conventional chocolate before the experiment (see Table 4 in Appendix D.2). Therefore, we conclude that the change in the classification trade-off is not problematic. Also the non-parametric Mann-Whitney tests for the two subsamples are consistent (see Appendix D.2).

but very small and far from being significantly different from zero ($p > 0.8$) when we control for earnings from the first part of the experiment.[23]

In line with Hypothesis 1, we further find that choosing fair trade chocolate before the experiment is associated with a significantly higher fair trade premium (+31.3 Cents, $p < 0.001$). Due to the asymmetric treatment effect, the difference in the fair trade premium virtually disappears in the public treatment, in contrast to Hypothesis 3 (see Table 7 in Appendix D.4 for separate regressions of the fair trade premium for the two treatments). Higher earnings in the first stage also increase the stated fair trade premium; the effect is marginally significant (coefficient of 0.007, i.e. +0.7 Cents per €1 higher income, $p = 0.056$). The interaction between first stage earnings and *public* is significantly negative with a coefficient of -0.012 ($p = 0.047$). This implies that in the public treatment, first stage earnings have no significant effect. We confirm this in a separate regression conditioning on the treatment being public, where the coefficient of first stage earnings is insignificant ($p > 0.2$, see Table 7 in Appendix D.4).

While the significance of earnings is not surprising, as it illustrates a simple income effect,[24] at a first glance, the irrelevance of income in the public treatment is. A possible explanation is that income from the market game was substantially higher if an individual was a consumer than if she was a worker or a firm. Individuals with high earnings from the market game might shy away from stating very large willingness-to-pay in the public treatment so that they do not reveal their fortunate role allocation as a consumer. In Table 8 in Appendix D.5, we show that indeed we obtain almost identical results if we control for being a consumer (instead of earnings) and the interaction with the public treatment.

Our main results remain qualitatively unchanged if we exclude 17 individuals with "no demand," i.e., individuals who state a willingness-to-pay of less than 2 cents for each of the two types of chocolate (see Table 2, column 2). The main difference is that

[23]Table 6 in Appendix D.3 shows results from regressions on the two sub-samples of individuals who chose conventional and fair trade chocolate, respectively. It is only for those who chose conventional chocolate that the treatment effect is significant. Again, considering the results separately for the subsamples who chose a bar of fair trade chocolate over one, respectively two, bars of conventional chocolate provides consistent findings (see Table 5 in Appendix D.2).

[24]Indeed in the private treatment both w_{fair} and w_{conv} increase with the income from the first part of the experiment.

the earnings from the first part no longer have a significant impact (and the interaction effect with the treatment dummy is significant only at the 10% level).[25]

It is unclear how interaction in the market game within a group can have impacted on the individually and independently stated fair trade premium other than possibly through the earnings. Nonetheless, to control for spill-over effects from the interaction in the market game, we run robustness checks with group-level random effects (see Table 10, columns 3 and 4, in Appendix D.7) and in addition also clustering standard errors on the group level (with 36 independent groups, see Table 10, columns 5 and 6, in Appendix D.7). The main results are unchanged.

5 Explaining the heterogeneous treatment effect

As discussed in the previous section, our results that, on average, fair trade choosers do not react to the treatment, but conventional choosers do, would be consistent with a negative correlation between individuals' intrinsic preferences for fair trade and their desire for a positive social image (compare Proposition 3). In this section, we discuss how this negative correlation between intrinsic motivation and image concerns can arise in a psychologically plausible way, and provide a tractable model. We then discuss alternative explanations and argue how they are inconsistent with our data.

5.1 An extended model with endogenous image concerns

A negative correlation between image concerns and intrinsic motivation is psychologically plausible because social-psychological research shows that people are more prone to actively manage the impressions that others have about them if they, correctly or incorrectly, perceive that the image that they send differs from the image that they would like to send or that they consider socially acceptable (Leary and Kowalski, 1990). Specifically, "people with low self-esteem may respond to social pressure (and act less con-

[25] One participant, a fair trade chooser in the public treatment, stated after the experiment that he accidentally swapped w_{fair} and w_{conv} (this is one of the two participants with a negative fair trade premium, as shown in the histogram on the right hand side of Figure 4 below). The aggregate data and statistical analysis reported in this paper use the original data, as entered, because some participants always make mistakes and it seems somewhat arbitrary to correct those that some participants report later to be mistakes. Nevertheless, we also performed robustness checks with the WTPs that the participant says he meant to enter. The only difference we observe is that the significance of the impact of the earnings from the first part is weaker in some of the regressions, but the impact of the earnings is not our concern.

sistently with their inner compass) than do people with high self-esteem" (MacDonald and Leary, 2012, p.363). We translate this into an economic model where the concern for one's social image decreases in the individual's self-image. This model provides a foundation for the heterogeneity of marginal utility of social image through the need to make up for low self-image.

Formally, the intuition that individuals may care more about their social image, if they perceive a discrepancy between their own type and the social ideal S, is expressed by the following utility function:

$$\begin{aligned} EU(w_{\text{fair}}, w_{\text{conv}}) &= \frac{1}{2}\left[\int_0^{w_{\text{conv}}} v_{\text{conv}}(p_{\text{conv}})dF_{\text{conv}}(p_{\text{conv}}) + \int_0^{w_{\text{fair}}} v_{\text{fair}}(p_{\text{fair}})dF_{\text{fair}}(p_{\text{fair}})\right] \\ &+ \gamma s(m) + (S - s(m))\lambda R(a) \end{aligned} \tag{8}$$

where $s(m)$ is the self-image (normalized to lie between 0 and the social ideal S which can be normalized to $S = 1$ without loss of generality), which (weakly) increases in own type and γ is the weight assigned to self-image.[26] Utility increases in self-image and in social image. However, the marginal utility of social image decreases with self-image. The optimal exaggeration Δ of the fair trade premium in the public treatment is then determined by the first-order condition

$$(S - s(m))\lambda \frac{\partial R(m + \Delta)}{\partial \Delta} = k\Delta. \tag{9}$$

It is easy to see that the utility-maximizing Δ is falling in $s(m)$ and hence in m and, thus, those who are more strongly intrinsically motivated will react less to the possibility of social image building, unless $\frac{\partial R}{\partial \Delta}$ is (strongly) increasing. However, the quadratic cost function still implies that the optimal Δ is positive as long as $\frac{\partial R}{\partial \Delta} > 0$, which will hold in any separating equilibrium. Therefore, we would also expect fair trade choosers to inflate their fair trade premium in the public treatment. If their self-image is substantially

[26]Note that we assume here that agents are certain of their own type and do not signal their type to themselves by their stated fair trade premium. In Appendix C, we show that the predictions are very similar if we assume that the person is not entirely sure about her own type, and the realized self-image depends on the stated fair trade premium. We also note that the model expressed in (8) collapses to our base model with homogeneous preferences if participant's self-image is unaffected by their fair trade premium, i.e., $s(m)$ is constant in m for all participants and $\rho = S - s(m)$. Even if participants care about their self-image ($\gamma > 0$), this would then not affect choices, because the utility from self-image would be independent of their choices, but the model is identical if we further set $\gamma = 0$.

larger than that of the conventional choosers (as is plausible), the predicted treatment effect would be much smaller than for the conventional choosers. Given the individual heterogeneity in the true m (which we cannot observe) and our between-subject design, this might explain why we detect essentially no treatment effect at all for the fair trade choosers.

5.2 Alternative explanations

There are several conceivable alternative explanations for our results. Suppose that a positive image is realized by revealing a positive fair trade premium above a certain threshold, but this image does not further improve with the size of the premium. Then, the fact that fair trade choosers do not increase their fair trade premium is not informative about their image concerns. There are several possible versions of such a model. First, one could think of some ideal concern for fair trade such that nobody would like to be perceived as a type with a higher willingness-to-pay, along the lines of the conformity model investigated by Bernheim (1994).[27] However, in this case, the fair trade premium should cluster at a specific positive level in the public treatment, which it does not (see Figure 4). The modified version of this model that is used in Andreoni and Bernheim (2009) to explain audience effects in dictator game giving, does not work here either because then nobody should actually have an intrinsic fair trade premium above the threshold that yields the best possible image, which is inconsistent with the distribution of the willingness-to-pay observed in the private treatment (see Figure 4).

A more complicated model with such an interior image threshold would allow for intrinsic premiums to be above this threshold. Assuming that image is derived from signaling one's intrinsic willingness to support fair trade, the image threshold can only take an interior value if image utility does not increase in the perceived type throughout but if it instead reaches some satiation point.[28] This would be interpreted as individuals not wanting to be perceived as a really low type but being indifferent with respect to being perceived as a reasonably high type and any type above.

[27] Non-signaling models of moral motivation formalize a similar intuition by deriving an ideal action, typically an interior value, from Kantian reasoning. The closer an individual's action is to this ideal or norm, the higher is her moral utility derived from it. Exceeding the ideal, however, is as bad as falling short of it. See for instance, Brekke et al. (2003).

[28] Explanations along these lines have on several occasions been suggested to us by readers and listeners, so we discuss this here in some detail.

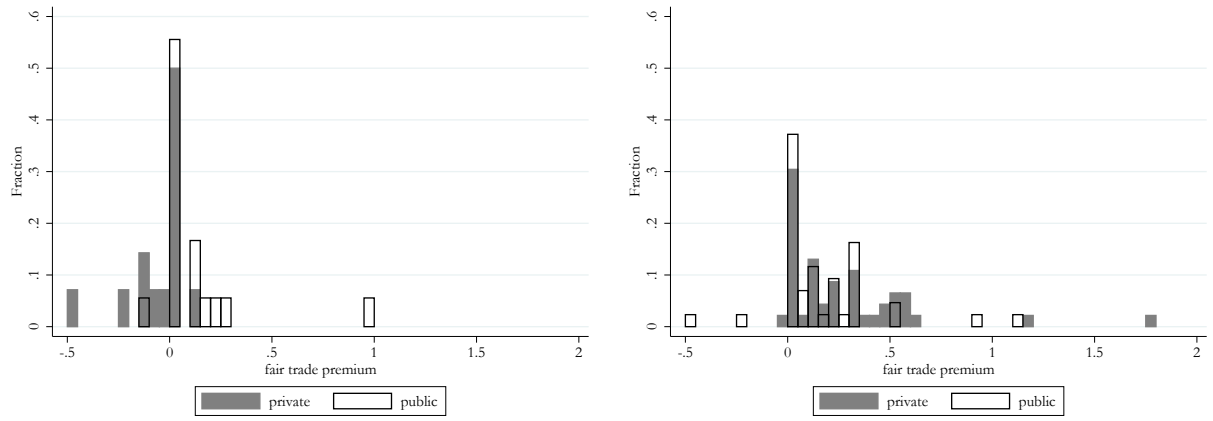

Figure 4: Histograms of distribution of fair trade premiums a, bin size: 5 cents. Left panel: conventional choosers, right panel: fair trade choosers.

We argue that such a model is also inconsistent with our data. Let us with some abuse of notation identify the types by the premium they choose in the private treatment because this should reflect their intrinsic motivation. Furthermore, denote the threshold type that people like to be (at least) perceived as by a_t, such that being perceived as a_t or any type $a > a_t$ yield the same image utility. Naively, one would think that types not too far below a_t would then choose a_t in the public treatment, inducing partial pooling at a_t. This, however, cannot happen in equilibrium, because such pooling would lead, upon observing a fair trade premium a_t, to the inference that the player is a type $a < a_t$ with positive probability. This implies that also an image-concerned type a_t would inflate his premium in public.[29] By the standard signaling logic, this creates pressure on the types above a_t to also increase their premiums in public. Now let us denote the equilibrium premium that type a_t chooses in public by $\bar{a}_t$. Then all premiums above $\bar{a}_t$ lead to the inference that the true type is some $a > a_t$. By assumption, being perceived as a better type than a_t yields the same utility as being perceived as a_t, and thus players with true types above $\bar{a}_t$ will not inflate their premiums in public beyond their intrinsic premium. In contrast, all those below $\bar{a}_t$ would inflate their premium in the public treatment but not beyond $\bar{a}_t$.

The observed absence of a treatment effect for the fair trade choosers would be consistent with this model if their intrinsic premium is larger than $\bar{a}_t$. This is conceivable. The intrinsic premium of conventional choosers should intuitively be smaller than a_t,

[29]Remember that the marginal cost of inflating one's premium is zero at zero inflation so that given a discrete image gain achieved by separating from the types $a < a_t$ would make this worthwhile for a_t.

which is the intrinsic premium beyond which types are classified as "sufficiently high". While conventional choosers should inflate their fair trade premium in the public treatment, they should not do so beyond $\bar{a}_t$. Hence, the distributions of fair trade premiums for the conventional and fair trade choosers should actually be disjoint. This would arguably be a very demanding test. Even if we, however, allow for some conventional choosers to have intrinsic premiums above $\bar{a}_t$, all those who inflate their premium in the public treatment and who are responsible for the significant increase in the fair trade premium of the conventional choosers should not inflate it above $\bar{a}_t$. Therefore, if such a model would capture both the lack of a treatment effect for the fair trade choosers and the treatment effect for the conventional choosers, the implied pattern in the public treatment would require that the stated fair trade premiums of the conventional choosers are still substantially below those of the fair trade choosers. This is not what we observe, since we do not detect statistically significant differences between the fair trade premiums of fair trade choosers and conventional choosers in the public treatment. Furthermore, looking at Figure 4, it is hard to make out where the threshold a_t and the resulting $\bar{a}_t$ should be.[30]

Another possible argument is that for fair trade choosers, the increase in image from separating compared to pooling with the conventional choosers could be small. In this case, the fair trade choosers would not find it worthwhile to increase their fair trade premium to achieve separation (for example, because relatively few conventional choosers would pool with them, which would, hence, not dilute their image much). This argument requires a perception of a dichotomous distribution of fair trade and conventional choosers; thus, the variance in observed positive fair trade premiums speaks against this hypothesis (see also Figure 4). Moreover, in such a model, the variance in fair trade premiums should be smaller in the public treatment, at least when focusing on the conventional choosers. This is not what we find. Using Levene's test, we cannot reject that

[30]One could also think that different participants have heterogenous beliefs about what would be the appropriate threshold for a good social image. In this sense, they could all be trying to pool, but because of non-equilibrium beliefs, they do not actually choose the same fair trade premium in the public treatment. Nevertheless, all participants who are still below their assumed threshold should inflate their fair trade premium in the public treatment. This is inconsistent with the absence of a treatment effect for the fair trade choosers unless they all perceive this threshold to be equal to their intrinsic fair trade premium. Furthermore, a model with so many degrees of freedom appears to be non-falsifiable.

the variances are equal across treatments, neither for the fair trade choosers ($p = 0.248$) nor for the conventional choosers ($p = 0.485$), nor for the pooled sample ($p = 0.154$).[31]

A further possible argument is that subjects might be motivated to build an image toward the experimenters. As a result, some of those whom we classify as intrinsically motivated might be rather concerned with their image in the eyes of the experimenter when replying to our recruitment email and they might also inflate their true fair trade premium in the private treatment. Note that such a kind of experimenter demand effect is not per se problematic for the question we are interested in. We are interested in who reacts to opportunities to build a social image. If subjects care about their image in the eyes of the experimenter, this would increase their willingness-to-pay in all treatments. Nevertheless, as long as they also care about their image in the eyes of the group of other participants, and this is what we typically understand by *social* image, our treatment variation would affect their behavior.[32] Thus, the absence of a treatment effect for the fair trade choosers would only be consistent with an interpretation that they specifically care about their image as perceived by the experimenter, but that increasing the audience to all participants does not strengthen their image concern. Conventional choosers, however, do react to the change in the audience. We are not aware of any model that could rationalize these heterogeneous responses to experimenter and audience in a parsimonious way.[33]

As we already pointed out, according to our post-experimental questionnaire, our participants on average consider fair trade products to be of higher quality. Therefore, one might wonder whether our participants who chose fair trade chocolate before the experiment just believe it to be of higher quality and whether quality perceptions also drive our treatment effect. This conjecture is not supported by our data. If we add the

[31] These results are unchanged (with larger p-values) if we exclude two outliers, namely the fair trade chooser with a very large fair trade premium in the private treatment and the conventional chooser with a fair trade premium of 1 in the public treatment.

[32] In terms of our model, if participants care about their image in the eye of the experimenter, λ would be positive already in the private treatment. We would argue, though, that it should still be larger in the public treatment.

[33] As an aside, since the economic benefits of fair trade are not convincingly established, it is unclear that the experimenter in an economics experiment would consider fair trade chocolate superior. Moreover, a relatively recent study does not find differences between behavior in a single-blind procedure where the experimenter is watching as compared to a double-blind procedure in the dictator game (Barmettler et al., 2012). If one is further worried about a demand effect in the sense that some subjects state a higher willingness-to-pay in the public treatment because they think we expect them to impress their fellow participants, one has to find a plausible story why intrinsic motivation for fair trade is (negatively) correlated with responding to this demand.

response to the question from our questionnaire about the quality advantages of fair trade products as a control (or this response as well as its interaction with the treatment dummy), our results on our proxy for fair trade preferences, *FTchoice*, and its interaction with the treatment dummy, *public*, are not changed qualitatively and the coefficients on quality perception and its interaction with the treatment dummy are insignificant (see Appendix D.6).

It is also conceivable that it is not the least concerned with fair trade who have the strongest desire to signal fair trade concerns, but rather those who are somewhat, but not strongly concerned about fair trade. These participants might believe that it is good in principle to buy fair trade products but are too selfish to pay much more for them. Nevertheless, they might want to signal strong concerns for fair trade. Such participants are likely to be classified as fair trade choosers when we offer the choice between one bar of conventional chocolate and one bar of fair trade chocolate (1-to-1) in the pre-experimental choice but as conventional choosers when we offer two bars of conventional chocolate (2-to-1). This conjecture would then imply that the treatment effect on conventional choosers should be larger and the treatment effect on the fair trade choosers should be smaller in the 2-to-1 sub-sample than in the 1-to-1 sub-sample. As we show in Appendix D.2, the (insignificant) differences between the two sub-samples point in the opposite direction with respect to both implications.

Finally, some participants might see a public good character in fair trade but do not care sufficiently about this public good to contribute in private. In public, however, they might decide to contribute and, thereby, be an example to others who might follow them and contribute in the future because then their costs of contribution would have an effect beyond that of their own consumption choice. Note that our experiment is one shot, meaning that such future contributions would have to happen outside the lab.

6 Discussion and implications

Our experimental results demonstrate that individuals may systematically differ in their image concerns, in contrast to the approach taken traditionally in the literature on image concerns and, in particular, conspicuous consumption. Therefore, it is important to account for this heterogeneity in both economic modeling and policy recommendations. Our results, suggesting a negative correlation between intrinsic motivation and image concerns, are in line with indirect evidence from several studies in other do-

mains. For instance, Filippin et al. (2013) analyze tax morale in Italian cities and find that those intrinsically motivated are less affected by the possibility of having a negative social reputation (for withholding taxes). This finding is corroborated by field experimental evidence from Germany (Boyer et al., 2016; Dwenger et al., 2016). Results by Riedl and Smeets (2017) indicate that among professional financial investors, intrinsic motivation and social image concerns are also negatively correlated. They show that "selfish" investors choose socially responsible mutual funds if these are not associated with tax benefits because they care about their image. Similar signaling motivations are not found among those who are classified as pro-social. Additionally, in the domain of conspicuous consumption, empirical evidence in Charles et al. (2009) is consistent with a negative correlation between wealth and the desire to signal wealth. In contrast, Grossman (2015), in a laboratory experiment, finds more compelling evidence for social signaling concerns if he excludes "selfish-types" and "money-maximizers," indicating a positive relation between image motivation and intrinsic giving in his sample, opposite to what we find.

While the accumulated evidence from the field is in line with the negative correlation we find, the absence of image concerns for the fair trade choosers in our experiment does not rule out that the same individuals exhibit concerns for their social image in other circumstances. However, we expect our findings to generalize in ethical consumption settings, where we would expect that the more intrinsically motivated consumers are less affected by image concerns than those consumers who care little intrinsically.

Our results have implications for the optimal design of policy interventions that intend to direct consumers toward more ethical purchasing behavior, as well as for the optimal design of marketing campaigns of private companies. Based on our findings, we expect that incentivizing consumers who are not sufficiently intrinsically motivated to buy ethically, for example by addressing their signaling desire, can increase ethical consumption without having to fear an image-based crowding out of intrinsically motivated buyers. Since the intrinsically motivated subjects in our experiment are not influenced by social image building opportunities, they would not be affected if the derived image is diluted because those not intrinsically motivated are encouraged by extrinsic incentives (such as image building opportunities or material rewards) to buy the same products. Hence, extrinsic incentives are not likely to crowd out intrinsic motivation in our setting. Indeed, crowding-out effects of pro-social behavior, to the best of our knowledge, are not observed in the context of ethical consumption, although

they are observed in non-market settings such as blood donations (Mellström and Johannesson, 2008; Lacetera and Macis, 2010). This suggests that the market for fair trade products can be enlarged by appealing to consumers' image concerns. However, the long run effect of a publicity campaign for fair trade would have to take into account possible market responses. The negative correlation between intrinsic motivation and image concerns would suggest that the profit maximizing strategy of a monopolist will attempt to pool consumers who intrinsically value fair trade with those who only care about their image as may in fact be a result of fair trade products being sold in specialty stores as well as discounters. In this case, increasing the value of the associated image through public campaigns may only increase prices without positive effects on either farmer or consumer welfare (see Friedrichsen, 2018, for the underlying theoretical results and further discussion).

We find that the elicited fair trade premiums and, thereby, total expected revenue for fair trade products increases with public exposure of individual decisions. The effect of increased public scrutiny on consumer welfare, however, depends on whether positive acts yield prestige or negative acts are stigmatized. In our study, the positive effect on the fair trade premium of the conventional choosers is mostly driven by an increase in their WTP for fair trade chocolate, in line with our model and the assumption that social prestige is derived from support for fair trade. This is in contrast to findings of Teyssier et al. (2015), who find an increase in the fair trade premium that is driven by a decrease in the WTP for conventional chocolate. One possible explanation for this difference is how the framing of the choice situation is perceived by the individuals. If it is perceived as attaching a positive image to fair trade, individuals will expect an image gain when purchasing the fair trade product and increase their WTP for the fair trade option. If it is perceived as attaching a negative image to conventional products, individuals will expect a utility loss from purchasing conventional chocolate and reduce their WTP for conventional chocolate accordingly. Cappelen et al. (2017) find evidence for both social esteem and social pressure in a dictator game study.

7 Conclusion

We address the heterogeneity in image concerns by studying the effect of opportunities for image building on fair trade premiums for experimental participants with different intrinsic motivation. We find that participants with low intrinsic motivation to buy fair

trade react positively to image building opportunities, while those with high intrinsic motivation do not. This is inconsistent with a standard model of image concerns that assumes identical image concerns for all participants. Instead, our results suggest that, in our setting, intrinsic motivation and image concerns are negatively correlated. We develop a model that captures this effect in a tractable way, based on a psychologically plausible interaction between self-image and social image. We argue that taking this correlation into account is important for consumer policy and firm behavior.

Acknowledgements

We thank Yves Breitmoser, Antonio Guarino, M'Hamed Helitim, Steffen Huck, Heiko Karle, Georg Kirchsteiger, Tobias König, Tatiana Kornienko, Johanna Möllerström, Rosemarie Nagel, Charlie Plott, Nora Szech, Mirko Tonin, Bertil Tungodden, Jean-Robert Tyran, Joël van der Weele, Christian Waibel, and Georg Weizsäcker as well as seminar participants at the University of Cologne, Humboldt-University Berlin, Paris I–Sorbonne, the University of Marburg, the University of Birmingham, the University of Nottingham, University of Vienna, Texas A&M, NYU, CAU Kiel, DIW Berlin, the Berlin Behavioral Economics workshop, and at the International ESA conference in New York, the European ESA conference in Cologne, the annual meeting of the GfeW in Karlsruhe, the HeiMaX workshop in Heidelberg, the Spring Meeting of Young Economists in Aarhus, the Journées Louis-André Gérard-Varet in Aix-en-Provence, the EEA in Gothenburg, the EARIE in Evora, the annual meeting of the Verein für Socialpolitik in Düsseldorf, the CESifo Area Conference on Behavioral Economics, the Nuremberg Experimental Research Days, the workshop on Consumer Behaviour, Self-Control and Intrinsic Motivation in Copenhagen, the Arne Ryde Workshop on Identity, Image and Economic Behavior, Image in Lund, and the workshop on Concerns for Status and Social Image at WZB Berlin for helpful comments and suggestions. We further thank Adam Lederer for language editing help. Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged. Jana Friedrichsen further acknowledges financial support by the Leibniz Competition through the project GlobalFood (SAW-2015-DIW-4).

References

Andreoni, J. and B. D. Bernheim (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77*(5), 1607–1636.

Ariely, D., A. Bracha, and S. Meier (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review 99*(1), 544–555.

Barmettler, F., E. Fehr, and C. Zehnder (2012). Big experimenter is watching you! Anonymity and prosocial behavior in the laboratory. *Games and Economic Behavior 75*(1), 17–34.

Becker, G., M. DeGroot, and J. Marschak (1964). Measuring utility by a single-response sequential method. *Behavioral Science 9*(3), 226–232.

Benabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies 70*(3), 489–520.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy 102*(5), 841–877.

Boyer, P., N. Dwenger, and J. Rincke (2016). Do taxes crowd out intrinsic motivation? Field-experimental evidence from Germany. *Journal of Public Economics 144*, 140–153.

Brekke, K. A., S. Kverndokk, and K. Nyborg (2003). An economic model of moral motivation. *Journal of Public Economics 87*(9), 1967–1983.

Cappelen, A. W., T. Halvorsen, E. Ø. Sørensen, and B. Tungodden (2017). Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association 15*(3), 540–557.

Carpenter, J. and C. K. Myers (2010). Why Volunteer? Evidence on the Role of Altruism, Image, and Incentives. *Journal of Public Economics 94*(11-12), 911–920.

Cason, T. N. and C. R. Plott (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy 122*(6), 1235–1270.

Charles, K. K., E. Hurst, and N. Roussanov (2009). Conspicuous consumption and race. *Quarterly Journal of Economics 124*(2), 425–467.

Dammert, A. C. and S. Mohan (2015). A survey of the economics of fair trade. *Journal of Economic Surveys 29*(5), 855–868.

Danz, D., D. Engelmann, and D. Kübler (2012). Do legal standards affect ethical concerns of consumers? An experiment on minimum wages. University of Mannheim Working Paper Series, No. 12-3.

De Pelsmacker, P., L. Driesen, and G. Rayp (2005). Do consumers care about ethics? willingness to pay for fair-trade coffee. *Journal of Consumer Affairs 39*(2), 363–385.

Dragusanu, R., D. Giovannucci, and N. Nunn (2014). The economics of fair trade. *Journal of Economic Perspectives 28*(3), 217–236.

Dwenger, N., H. Kleven, I. Rasul, and J. Rincke (2016). Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany. *American Economic Journal: Economic Policy 8*(3), 203–232.

Engelmann, D., J. Friedrichsen, and D. Kübler (2018). Fairness in markets and market experiments. Collaborative Research Center TRR 190 Rationality and Competition, Discussion Paper No. 64.

Filippin, A., C. V. Fiorio, and E. Viviano (2013). The effect of tax enforcement on tax morale. *European Journal of Political Economy 32*, 320–331.

Fischbacher, U. (2007). Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Friedrichsen, J. (2018). Signals sell: Product lines when consumers differ both in taste for quality and image concern. Collaborative Research Center TRR 190 Rationality and Competition, Discussion Paper No. 70.

Glazer, A. and K. A. Konrad (1996). A signaling explanation for charity. *American Economic Review 86*(4), 1019–1028.

Gneezy, U. and A. Rustichini (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics 115*(3), 791–810.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125.

Griskevicius, V., J. M. Tybur, and B. Van den Bergh (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology 98*(3), 392–404.

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization 117*, 26–39.

Harbaugh, W. T. (1998). What do donations buy? A model of philanthropy based on prestige and warm glow. *Journal of Public Economics 67*(2), 269–284.

Hillman, A. L. (2010). Expressive behavior in economics and politics. *European Journal of Political Economy 26*(4), 403–418.

Hopkins, E. and T. Kornienko (2004). Running to keep in the same place: Consumer choice as a game of status. *American Economic Review 94*(4), 1085–1107.

Koppel, H. and G. G. Schulze (2013). The importance of the indirect transfer mechanism for consumer willingness to pay for fair trade products: Evidence from a natural field experiment. *Journal of Consumer Policy 36*(4), 369–387.

Lacetera, N. and M. Macis (2010). Do all material incentives for pro-social activities backfire? The response to cash and non-cash incentives for blood donations. *Journal of Economic Psychology 31*(4), 738 – 748.

Leary, M. R. and R. M. Kowalski (1990). Impression management: A literature review and two-component model. *Psychological Bulletin 107*(1), 34.

Loureiro, M. L. and J. Lotade (2005). Do fair trade and eco-labels in coffee wake up the consumer conscience? *Ecological Economics 53*(1), 129–138.

MacDonald, G. and M. R. Leary (2012). Individual differences in self-esteem. In M. R. Leary and J. P. Tangney (Eds.), *Handbook of Self and Identity*, Chapter 17, pp. 354–377. Guilford Press.

Mellström, C. and M. Johannesson (2008). Crowding out in blood donation: Was Titmuss right? *Journal of the European Economic Association 6*(4), 845–863.

Miller, G. F. (2009). *Spent: Sex, Evolution, and Consumer Behavior*. Viking Adult.

Munro, A. and M. Valente (2016). Green goods: Are they good or bad news for the environment? evidence from a laboratory experiment on impure public goods. *Environmental and Resource Economics 65*(2), 317–335.

Plott, C. R. and K. Zeiler (2005). The willingness to pay/willingness to accept gap, the "endowment effect", subject misconceptions and experimental procedures for eliciting valuations. *American Economic Review 95*(3), 530–545.

Read, D., G. Loewenstein, and M. Rabin (1999). Choice bracketing. *Journal of Risk and Uncertainty 19*(1-3), 171–197.

Reinstein, D. and J. Song (2012). Efficient consumer altruism and fair trade products. *Journal of Economics & Management Strategy 21*(1), 213–241.

Riedl, A. and P. Smeets (2017). Why do investors hold socially responsible mutual funds? *The Journal of Finance 72*(6), 2505–2550.

Sexton, S. E. and A. L. Sexton (2014). Conspicuous conservation: The Prius halo and willingness to pay for environmental bona fides. *Journal of Environmental Economics and Management 67*(3), 303–317.

Soetevent, A. R. (2011). Payment choice, image motivation and contributions to charity: Evidence from a field experiment. *American Economic Journal: Economic Policy 3*(1), 180–205.

Teyssier, S., F. Etilé, and P. Combris (2015). Social-and self-image concerns in fair-trade consumption. *European Review of Agricultural Economics 42*(4), 579–606.

A Proofs

Proof of Proposition 1. If an agent chooses her WTP such that the observed fair trade premium is a and the WTP for the conventional bar of chocolate is u, an observer infers that for this agent the following holds:

$$a = m + \frac{\rho}{k}\frac{\partial R(m+\Delta)}{\partial \Delta} + \varepsilon \tag{10}$$

Knowing that stated values may differ from true ones only by a mean-zero, normally distributed noise term, the observer expects that this agent has a fair trade preference equal to

$$E[m|a] = a - \frac{\rho}{k}\frac{\partial R(m+\Delta)}{\partial \Delta} \tag{11}$$

We use the observation that $\frac{\partial R(m+\Delta)}{\partial \Delta} = \frac{\partial R(a)}{\partial a} = R'(a)$, and that the social image is defined as the conditional expectation of an individual's fair trade preference m upon observation of her fair trade premium and WTP for the conventional chocolate, $R(a) = E[m|a]$. Then, the public inference constitutes a first-order differential equation

$$R(a) = a - \frac{\rho}{k}R'(a) \tag{12}$$

for which the generic solution is $R(a) = a - \frac{\rho}{k} + e^{-ak/\rho}C$ for a constant C. The agent's optimization problem is globally concave only for $C = 0$, so that the unique solution for our context is

$$R(a) = a - \frac{\rho}{k} \tag{13}$$

Hence $R'(a) = 1$ and $a = m + \frac{\rho}{k}$ by (10).

□

The argument in the proof of Proposition 1 relies on the fact that the base utility for chocolate u is observed. The inference is calculated as if it is done separately for any possible u, because, in principle, in equilibrium the stated fair trade premium could depend on u, which is also observed. The result, however, is independent of u and, hence, the same inflation of image is obtained for any base utility of chocolate, u, and any level of intrinsic motivation, m.[34]

Proof of Proposition 2. Assume that $R(\hat{a}) = \hat{a}$, i.e., upon observing a stated fair trade premium the public infers this to be the individual's attitude toward fair trade. Such naive inferences would be accurate if applied to WTP that were stated in private or if individuals' WTP where unaffected by public exposure. The term on the left-hand side of (6) is

[34]Note that in the experiment, participants could not influence what type of chocolate they would obtain, but only whether they would obtain the randomly selected type. Hence, the trade-off they make is between the price they would pay for the fair trade chocolate and the image they might gain as a result.

positive for every $a = m + \Delta$, if $\rho > 0$. Thus, we would expect individuals of all types who care about social image to increase their stated WTP for fair trade in response to public exposure. Thus, the distribution of WTP shifts upwards. Moreover, the marginal gain in reputation is constant for naive inferences so that all WTP are shifted upward by the same amount. □

Proof of Proposition 3. Consider first the case that the conventional choosers (those with lower m) have systematically lower concerns for social image ρ_{conv} than the fair trade choosers ρ_{fair}. Assume initially that the derivative of the inference function R' is again constant. This implies that the fair trade choosers will inflate their fair trade premium by Δ_{fair} such that $\rho_{\text{fair}} R'(m_{\text{fair}} + \Delta_{\text{fair}}) = k\Delta_{\text{fair}}$ and conventional choosers by Δ_{conv} such that $\rho_{\text{conv}} R'(m_{\text{conv}} + \Delta_{\text{conv}}) = k\Delta_{\text{conv}}$. For R' constant this would imply that $\Delta_{\text{fair}} > \Delta_{\text{conv}}$. Then, however, because fair trade choosers have a higher m, the equilibrium inference would have to discount high fair trade premiums more, such that R is concave and not linear as assumed. However, the concavity of R cannot be so extreme that $\Delta_{\text{fair}} \leq \Delta_{\text{conv}}$ because then the equilibrium inference R would have to take this into account and not be concave. Thus in equilibrium if fair trade choosers care more about social image than conventional choosers, they will inflate their fair trade premium more than conventional choosers, but the difference will be smaller than if R' was constant. The same logic applies to the case that conventional choosers have systematically higher concerns for social image. Then the equilibrium inference R has to be convex, but not to a degree such that $\Delta_{\text{conv}} \leq \Delta_{\text{fair}}$ and as a result conventional choosers will inflate their fair trade premium more than fair trade choosers but again the difference will be smaller than it would be if equilibrium inference R' was constant. □

B Test for size effects

As argued in the main text, Hypothesis 1 would not only result if fair trade choosers cared more for fair trade, but also if they cared less for chocolate. Specifically, the observed fair trade premium should decrease in the chocolate valuation because $a = (x - d)u + m - xu + \varepsilon = m - du + \varepsilon$. We therefore consider the following auxiliary hypothesis, which is derived from the type classification.

Hypothesis 5. *(Size effects) Individuals who chose the fair trade chocolate have on average lower valuations for chocolate as such, i.e.* $E[w_{conv}(u, m)|\textit{chose fair}] < E[w_{conv}(u, m)|\textit{chose conv}]$.

In contrast to Hypothesis 5, the difference in fair trade premiums is not in part driven by lower willingness-to-pay for conventional chocolate by the fair trade choosers. Regressing the chocolate choice before the experiment on the stated willingness-to-pay for conventional chocolate, the latter is insignificant ($p = 0.895$ in private, $p = 0.979$ in public, and $p = 0.919$ pooled). Additionally, non-parametric tests yield no evidence that those with lower valuation for chocolate are more likely to choose fair trade chocolate, but instead the sorting is driven by fair trade preferences. In the private treatment,

the WTP for the conventional chocolate bar is nearly identical across the two groups ($w_{\text{conv}} = 0.45$ for fair trade choosers and $w_{\text{conv}} = 0.46$ for conventional choosers, Mann-Whitney test $p = 0.467$), whereas the WTP for the fair trade chocolate bar is substantially higher for fair trade choosers ($w_{\text{fair}} = 0.71$) than for conventional choosers ($w_{\text{fair}} = 0.40$), (Mann-Whitney test, $p = 0.051$; $p = 0.022$ if we exclude participants with WTP below 2 cents for both bars of chocolate).

C Extension to imperfect self-knowledge

Here we show that the conclusion from the self-knowledge case presented in Section 5.1 remains true if we assume that individuals are unsure of their type. In particular, we allow the self-image to positively depend on the stated fair trade premium.

Then, utility is given by the following equation

$$EU(w_{\text{fair}}, w_{\text{conv}}) = \frac{1}{2}\left[\int_0^{w_{\text{conv}}} v_{\text{conv}}(p_{\text{conv}})dF_{\text{conv}}(p_{\text{conv}}) + \int_0^{w_{\text{fair}}} v_{\text{fair}}(p_{\text{fair}})dF_{\text{fair}}(p_{\text{fair}})\right] + \gamma s(m,a) + (S - s(m,a))\lambda R(a) \quad (14)$$

where $s(m,a)$ is the self-image (normalized to lie between 0 and S), which (weakly) increases in own type and also increases in the stated fair trade premium.

Utility increases in self-image and in social image. However, the marginal utility of social image is decreasing in self-image. The optimal exaggeration of the fair trade premium in the public treatment is then determined by the first-order condition

$$\gamma\frac{\partial s(m, m+\Delta)}{\partial \Delta} - \lambda\frac{\partial s(m, m+\Delta)}{\partial \Delta}R(m+\Delta) + (S - s(m, m+\Delta))\lambda\frac{\partial R(m+\Delta)}{\partial \Delta} = k\Delta. \quad (15)$$

As the self-image is imperfect, an increase in the stated fair trade premium improves the individual's self image, $\partial s(m, m+\Delta)/\partial\Delta \geq 0$. Then, for any given level of motivation m, the first term on the left-hand side of (15) is positive and represents the direct effect of an improved self-image on utility. This effect is also present in the private elicitation though, so that it would not count into our treatment effect. The second term is negative and represents the decrease in the utility from social image due to increases in self-image. The third term is positive but decreasing in m and captures that marginal utility from social image is smaller for larger self-image. The total effect of an increase in the stated fair trade premium on utility depends on the relative importance of self- and social signaling motivations.

In light of our experimental design, we are mostly interested in the effect of making choices public. To that respect, the extension to include uncertainty about one's own type does not qualitatively affect our predictions. Assuming that both s and R have constant slopes, it becomes apparent that individuals with a higher level of intrinsic

motivation will react less strongly to the public treatment for two reasons. As in the case with self-knowledge, individuals with higher levels of intrinsic motivation have a lower marginal utility from the social image (see the third term in (15)). In addition, increasing the stated fair trade premium in response to public observation improves an individual's self-image and, thus, lowers the utility from the social image (see the second term in (15)). This marginal loss in utility is higher for individuals with higher intrinsic motivation because they expect (and realize) a higher social image.

If the loss in marginal utility of social image that is induced by inflating the stated WTP differential and the associated increase in the self-image is larger than the marginal increase in utility from social image (i.e., if the second term in (15) is absolutely larger than the third term), then an individual may even state a lower fair trade premium in public than in private. This is more likely to be the case the higher is the intrinsic preference for fair trade because then both the self-image and the social image are high. Such an effect would, however, be inconsistent with a separating equilibrium and the standard assumption of a differentiable image function for a continuous type distribution.