ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Fries, Tilman

Working Paper

Replication Report: The Welfare Effects of Pride and Shame

I4R Discussion Paper Series, No. 64

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Fries, Tilman (2023) : Replication Report: The Welfare Effects of Pride and Shame, I4R Discussion Paper Series, No. 64, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/276252

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

INSTITUTE for REPLICATION

No. 64
I4R DISCUSSION PAPER SERIES

Replication Report: The Welfare Effects of Pride and Shame

Tilman Fries

September 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 64

Replication Report: The Welfare Effects of Pride and Shame

Tilman Fries[1]

[1]*WZB, Berlin/Germany*

SEPTEMBER 2023

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the ZBW – Leibniz Information Centre for Economics, and RWI – Leibniz Institute for Economic Research, I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Editors

Abel Brodeur
University of Ottawa

Anna Dreber
Stockholm School of Economics

Jörg Ankel-Peters
RWI – Leibniz Institute for Economic Research

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research

Hohenzollernstraße 1-3
45128 Essen/Germany

www.i4replication.org

ISSN: 2752-1931

Replication Report: The Welfare Effects of Pride and Shame

Tilman Fries*

July 26, 2023

Abstract

This replication report examines and extends the research conducted by Butera, Metcalfe, Morrison, and Taubinsky (2022) on "The Welfare Effects of Pride and Shame." The original paper explores the welfare implications of public recognition as a motivator for desirable behavior and introduces an empirical methodology to measure Public Recognition Utility (PRU), which quantifies the utility individuals experience when their actions are publicly recognized. This report focuses on the real effort experiment reported in the paper that was conducted using a classroom sample, a lab sample, and an online sample. I computationally reproduce the original results and verify their robustness. While reproducing the results, I found two minor coding errors in the replication package. Correcting these errors slightly changes some estimates reported in the paper but does not turn over any results. The main treatment effect findings are further robust to using different sets of controls and sample selection criteria. Moreover, I conduct a heterogeneity analysis which reveals significant variations in how participants value public recognition. Overall, the replication study confirms the original conclusions while providing additional insights into the heterogeneity of PRU shapes on an individual level.

*WZB Berlin: tilman.fries@wzb.eu

1 *Introduction*

This report replicates and extends the research conducted by Butera et al. (2022)—or BMM&T—on "The Welfare Effects of Pride and Shame." Published in the *AER* in 2022, the paper explores the welfare implications of public recognition in motivating desirable behaviors. BMM&T introduce an empirical methodology to experimentally identify *Public Recognition Utility* (PRU), i.e., the utility that individuals experience when their actions are publicly recognized by others. This allows the authors to empirically quantify the welfare effects of public recognition interventions. In this replication study, I seek to assess the reliability of the original findings while delving into additional avenues of analysis and understanding.

Public recognition interventions are commonly used as a motivational tool in various domains, and a large literature exists that studies their behavioral effects (e.g., Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009; Bursztyn and Jensen, 2017). However, empirical measurements of the welfare effects of these policies are scarce. BMM&T focus on two issues related to welfare. First, they point out that the shape of the PRU will determine whether a public recognition intervention will increase or decrease average utility or keep it constant. Second, they clarify how the behavioral effect of a public recognition intervention that is scaled up to the whole population might be different from the reduced form effect that can be identified through an experiment. This can happen because, in many social image models that are used to model the consequences of public recognition, actions are evaluated relative to a moral standard. For example, the standard may be equal to the average action in the population. Scaling up an intervention to the whole population has a direct motivational effect as individual wants to appear "good" in front of others and an indirect effect, as the social standard—what individuals perceive to be "good behavior"—changes if everyone receives the treatment. BMM&T address both issues by combining experimental data with structural modeling.

Identifying the shape of the PRU and predicting the general equilibrium effects of public recognition requires structural measurements of the PRU. In a series of incentivized experiments, BMM&T elicit participants' willingness to pay (WTP) to have a socially desirable action of level x revealed to others. Within individuals, they vary x, which allows them to derive an estimate of WTP as a function of x. The paper then uses these estimates to back out the PRU and investigate welfare. The paper's results suggest that, while public recognition increases socially desirable behavior in all samples, the welfare effects are more ambivalent; across different experimental samples, public recognition either increases or decreases welfare or does not cause measurable differences.

In this replication report, I focus on a number of real effort experiments reported in the paper. These experiments were conducted using a classroom sample from Boston University, a student lab sample from Berkeley, and an online convenience sample recruited via Prolific. In these experiments, participants perform in a task earning points, and, across three within-participant treatments, their task points are either converted into charity donations or into charity donations and individual payments. In two treatments (one charity-only and one combining charity donations with individual payments), individual points are not disclosed to others. A third treatment combines donation-only with public recognition. Here, there is a chance that the individual point score will be disclosed to other participants in the experiment. Whether the public recognition treatment actually reveals a participant's points depends on their WTP statements: Before engaging in the real effort task, participants answer how much they would be willing to pay or accept to have their points revealed if their point score is within an interval

x. Participants answer 18 questions of this format which gradually increase the interval boundaries. After the experiment, an incentive-compatible mechanism determines whether the participant's actual task points will be revealed to others.

The primary objectives of this replication report are twofold: first, I computationally reproduce the original results reported in the paper and investigate the robustness of the main treatment effects with a specification curve analysis. Second, I conduct a heterogeneity analysis that investigates PRU shapes on the individual level.

I can computationally reproduce all results reported in the paper using the paper's replication package. Looking into the code, I find two minor coding errors and correcting them changes some of the estimates reported in the paper. For example, correcting an error that led the authors to improperly control for the order in which participants went through the different treatments suggests that, contrary to what is reported in the paper, order had a significant effect on behavior in two out of the three samples. Correcting the errors does not, however, change any of the paper's conclusions. I also show that the reported treatment effects are robust to using different subsets of controls and a different criterion to identify "incoherent" participant responses.

The heterogeneity analysis suggests that there is some heterogeneity in how participants value public recognition within the different experimental samples. Comparing the distribution of different PRUs across samples, I find that the online sample is least sensitive to public recognition and the classroom sample is most sensitive. I also show that there is a correlation between the individually perceived moral standard and the shape of individual PRU functions.

2 *Computational replication*

The original paper comes with a well-documented replication package. The only minor obstacle I faced when reproducing the paper's results was in the readme, which specifies that the provided code can be executed with Stata 15. Parts of the code, however, use syntax that was only introduced in Stata 16. The code did compile fine when using Stata 17 and produced all tables and graphs that were reported in the paper's main body and appendices.

While reproducing the paper's results, I found two errors in the original code. The first error was in the coding of the order dummies that are included in regressions reported in Table 5 of the paper. The table's regressions use data on the participant-treatment level; since the treatments are within-participant, there are three observations for every participant, one for every treatment. The authors thus control for order effects. However, their code included an error when assigning different order values to different observations. Appendix A explains the error in more detail.

Table 1 presents results of the main treatment effects using a corrected and the original coding of the order variable. The point estimates of the treatment dummies modestly increase by roughly 10% in the BU sample but otherwise remain similar to those that were reported in the original paper. They also remain statistically significant at the same levels. The point estimates of the order dummies increase for all samples and they become jointly significant at the 10% level in the Berkeley and BU samples.

Table 1: Replication of Table 5 with original and corrected order variable coding

	Table 5 column (1)		Table 5 column (2)		Table 5 column (3)		Table 5 column (4)	
	Original	Corrected	Original	Corrected	Original	Corrected	Original	Corrected
Public Recognition	105.0***	104.0***	134.4***	134.7***	103.6**	115.8***	106.7***	104.8***
	(12.25)	(12.27)	(22.56)	(22.53)	(45.25)	(43.93)	(18.72)	(18.61)
Financial Incentives	185.7***	186.4***	177.8***	177.1***	118.3***	132.3***	192.0***	193.3***
	(12.56)	(12.65)	(22.04)	(22.02)	(39.62)	(39.15)	(18.98)	(19.16)
Order=2	-13.93	-1.697	-22.00	31.48	-40.14	96.26**	-14.00	-1.767
	(11.95)	(11.92)	(21.42)	(21.30)	(39.39)	(42.58)	(11.96)	(11.92)
Order=3	-22.18*	-14.47	-20.24	66.92***	-89.33**	51.77	-22.22*	-14.89
	(12.40)	(13.74)	(20.84)	(23.66)	(42.95)	(44.33)	(12.41)	(13.72)
Group of 300							20.61	20.32
							(39.85)	(39.84)
Group of 300 × Public Recognition							-3.117	-1.203
							(28.43)	(28.41)
Group of 300 × Financial Incentives							-18.85	-19.89
							(29.05)	(28.98)
Group of 15							17.70	17.67
							(41.13)	(41.09)
Group of 15 × Public Recognition							-3.213	-1.815
							(31.13)	(31.06)
Group of 15 × Financial Incentives							-3.270	-4.590
							(31.90)	(31.91)
Control mean	807.921	807.921	989.755	989.755	815.924	815.924	807.921	807.921
Order dummies F-test	.18	.497	.497	.018	.116	.081	.178	.476
Sample	Prolific	Prolific	Berkeley	Berkeley	BU	BU	Prolific	Prolific
Observations	2904	2904	1152	1152	354	354	2904	2904

Note: The dependent variable in all regressions is points scored. The original Table 5 suppresses the output of the order dummies but they are printed here to facilitate comparison between the original and the corrected estimates. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2: Replication of Table 6 with original and corrected curvature calculation

	Table 6 column (2)		Table 6 column (4)		Table 6 column (6)	
	Original	Corrected	Original	Corrected	Original	Corrected
Points (100s)	0.155***	0.155***	0.379***	0.379***	0.309***	0.309***
	(0.0177)	(0.0177)	(0.0702)	(0.0702)	(0.116)	(0.116)
Points (100s) sqd.	-0.00365***	-0.00365***	-0.00404	-0.00404	0.00229	0.00229
	(0.000805)	(0.000805)	(0.00356)	(0.00356)	(0.00574)	(0.00574)
Constant	-0.733***	-0.733***	-3.325***	-3.325***	-5.076***	-5.076***
	(0.121)	(0.121)	(0.420)	(0.420)	(0.810)	(0.810)
$-R''/R'(\bar{a}_{pop})$	0.076	0.076	0.026	0.026	-0.013	-0.013
95 percent CI	[0.047, 0.106]	[0.047, 0.106]	[-0.018, 0.070]	[-0.018, 0.070]	[-0.079, 0.053]	[-0.079, 0.053]
$-R''/R'(\bar{a}_{pop}) \times SD$	0.245	0.394	0.110	0.135	-0.077	-0.067
95 percent CI	[0.186, 0.303]	[0.244, 0.549]	[-0.046, 0.267]	[-0.093, 0.363]	[-0.505, 0.351]	[-0.409, 0.275]
Sample	Prolific	Prolific	Berkeley	Berkeley	BU	BU
N	16456	16456	6528	6528	2006	2006

Note: Differences between the original and corrected estimates are highlighted in red. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Changing the results of Table 5 also impacts welfare estimates reported in tables 7 and 9 of the paper, as the estimated treatment coefficients are inputs for the structural estimation. Appendix C reports updated variants of these tables. However, since the consequences of this error for the estimated treatment effects are relatively minor, the welfare estimates barely change.

The second coding error that I found concerns the results presented in Table 6 of the original paper. This table presents estimates of the WTP for public recognition as a function of effort in the charitable contribution experiments. Of special interest in that table are estimates of the curvature of the PRU function. The code that calculated the unitless curvature measure, $-R''/R'(\bar{a}_{pop}) \times$ SD, contained a small mistake. Appendix B displays the original code and the adjustments that I made.

Table 5 compares the original Table 6 estimates with the corrected estimates. The error consequences are greatest for the Prolific sample, where the estimated curvature coefficient increases from 0.245 to 0.394 and the original point estimate is barely included in the corrected 95% confidence interval. This makes the unitless curvature point estimate of the Prolific sample the largest found in the paper, and places the point estimates of the curvature found for the YMCA sample (reported in tables 3 and 4 of the original paper) safely between those of the Prolific and Berkeley samples. The paper's original claim that point estimates for Prolific and YMCA are similar thus should be reconsidered in light of the new findings. However, also after correcting the coding error, the curvature estimate for the Prolific sample remains insignificantly different from that of the YMCA sample.

3 *Robustness and extensions*

This part presents a robustness analysis and an extension. I investigate the robustness of the treatment effects reported in Table 5 using different sample selection criteria and sets of controls. I additionally investigate the heterogeneity of the elicited PRUs within the different experimental samples.

3.1 *Specification curve analysis*

Table 5 in the original paper displays treatment effect estimates of increased public recognition or financial incentives on real effort. I conducted a specification curve analysis to investigate the robustness of the treatment effects to using different sample selection criteria and sets of controls.

Sample selection. The analysis in the paper excludes data from participants who fulfill at least one of the following criteria:

- They fail an attention check.
- They attain a very high points score in at least one of the rounds of the experiment.
- They report "incoherent" preferences for public recognition.

The original analysis classifies a participant as coherent (and thus does not exclude them) if (i) the sign of the stated WTP for public recognition switches not more than once when gradually increasing the points interval from *0-100* to *1600 -1700*, or, if (ii) the stated WTP for public recognition first switches its sign from negative to positive and thereafter from positive to negative. Therefore, WTP responses are deemed to be coherent if they go negative-positive, positive-negative, or negative-positive-negative.

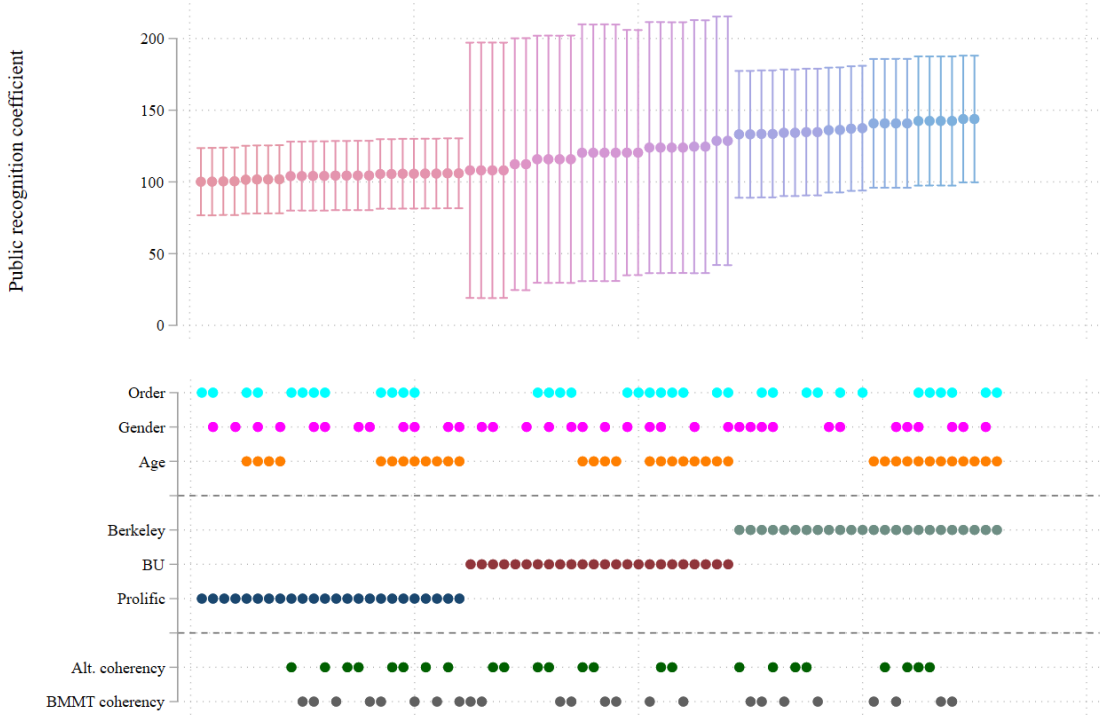
I propose an alternative coherency criterion which classifies a participant as coherent if the sign of the *difference* between the WTP statements of two adjacent intervals switches no more than once. To understand how this definition is different, consider the WTP statements of a (hypothetical) Participant A displayed in the table below. BMM&T's definition would classify this participant as coherent, as the stated WTP always remains negative. However, Participant A would not be classified as coherent under the alternative coherency measure; gradually increasing the points interval, the WTP first increases, then decreases, and then increases again. Similarly, Participant B would be classified as incoherent by the original paper's criterion but not by mine as their WTP is inverse-U shaped. Participant C on the other hand states an increasing WTP, which is coherent under both criteria.

Table 3: Examples of hypothetical participants and whether they would be classified as coherent

Points interval	Participant A	Participant B	Participant C
0-100	-10	10	-10
101-200	-2	-4	-2
201-300	-8	12	5
301-400	-1	14	7
...	...	...	...
Classified as coherent?			
BMM&T coherency	Yes	No	Yes
Alt. criterion	No	Yes	Yes

The alternative criterion therefore does not only consider the sign of the WTP. Instead, it asks whether the participant becomes more or less willing to pay for public recognition as their point score increases. The alternative criterion classifies 164 participants as incoherent in the Prolific sample, 48 in the Berkeley sample, and 20 in the BU sample. These numbers are 42, 11, and 2 under the coherency criterion used in the paper.

Figure 1: Specification curve plot of the public recognition coefficient



Note: The figure displays the estimated treatment effect of public recognition as a function of control variables and sample selection criteria. A dot indicates that a certain control variable was included in the regression or that a certain sample selection criterion was used. Order and gender are categorical variables, while age entered the regression in a continuous and linear way. Error bars are 95% confidence intervals.

Control variables. The specifications reported in the paper control for order effects. When controlling for order in the specification curve analysis below, I will use the corrected coding for the order variable, as explained in Section 2. In addition, the experiment elicited participants' age and gender, which were not used in the analysis results reported in the paper. To get a sense about how sensitive the estimated treatment coefficients are to including demographic variables and controlling for potential order effects, I will include them in the robustness analysis

Figure 1 displays a specification curve graph for the public recognition coefficient under different regression specifications, coherency criteria, and experimental samples. Appendix C displays the specification curve for the financial stakes coefficient. Overall, the treatment coefficient estimates are robust the different coherency criteria and the inclusion of the order and demographic variables. None of the considered specifications would change the ranking of the public recognition coefficient between the three experimental samples.

3.2 *Heterogeneity analysis*

The experiments reported in the paper suggest that there is substantial heterogeneity in the shape of the public recognition utility function (PRU) between experimental samples (Berkeley, BU, and Prolific). BMM&T assume that the shape of the PRU as a function of

the point score a takes on the following parametric form:

$$PRU(a) = \beta_0 + \beta_1 a + \beta_2 a^2. \quad (1)$$

If the PRU increases in a, it can be characterized by a curvature parameter which denotes whether the function is convex ($\beta_2 > 0$) or concave ($\beta_2 < 0$) and a parameter denoting the moral standard. The moral standard, s solves the equation $PRU(s) = 0$, essentially denoting the point score at which the individual experiences zero PRU from disclosing their effort choice. These two parameters both will influence whether, in utilitarian welfare terms, public recognition interventions will be desirable or undesirable. As the moral standard increases, more individuals will receive negative utility from public recognition, making it less desirable. Similarly, as the PRU becomes more concave, undesirable actions will be more stigmatized while desirable actions will be less honored. Therefore, a more concave PRU will tend to decrease welfare (see the original paper by BMM&T for a more detailed and formal discussion).

When estimating the PRU for each sample, BMM&T aggregate the WTP statements from all sample participants in one analysis, essentially assuming that all sample participants have the same (average) utility function. This is a strong, though common and defensible assumption. In order to shed some light on the underlying utility functions that are being aggregated, I conduct an additional analysis for heterogeneity of the PRU within these samples. Because the experiment uses a strategy method design to elicit the PRU for different levels of a, we can estimate Equation (1) at the individual level.[1]

For each individual in the data set, I estimate the parameters of Equation (1) at the individual level.[2] I classify participants into the following six categories.

- *Incoherent:* Using BMM&T's coherency criterion.
- *Insensitive:* Individuals state the same WTP for all point score intervals.
- *Inverse-U:* $\hat{PRU}(a)$ is estimated to be convex with a minimum at $a \in (50, 1650)$.
- *Hump-shaped:* $\hat{PRU}(a)$ is estimated to be concave with a maximum at $a \in (50, 1650)$.
- *Decreasing:* $\hat{PRU}(a)$ is either estimated to be convex with a minimum at $a \geq 1650$ or is estimated to be concave with a maximum at $a \leq 50$.
- *Increasing:* $\hat{PRU}(a)$ is either estimated to be convex with a minimum at $a \leq 50$ or is estimated to be concave with a maximum at $a \geq 1650$.

Figure 2 shows that there is a substantial variation of types within each sample. Consistent with what one might expect, online participants on Prolific are least sensitive to changes in disclosed points, while classroom participants from BU are most sensitive.

[1] BMM&T report estimates of a mixed-effects model in their Online Appendix Section E where the coefficients of Equation (1) are assumed to be jointly normally distributed and to vary on the individual level. They find little variation in the PRU curvature across individuals but more variation in the moral standard. In comparison to their approach, my approach below allows for more flexibility by not putting any structure on how the coefficients are distributed in the population. However, a downside of this is that my approach might extrapolate too much from noisy responses to the WTP questions.

[2] While the experiment elicited WTP for different point *intervals*, Equation 1 is a function of point *levels*. I follow the methodology used by BMM&T and replace the interval with its midpoint in the estimation.

Figure 2: Estimated PRU type distribution, by sample

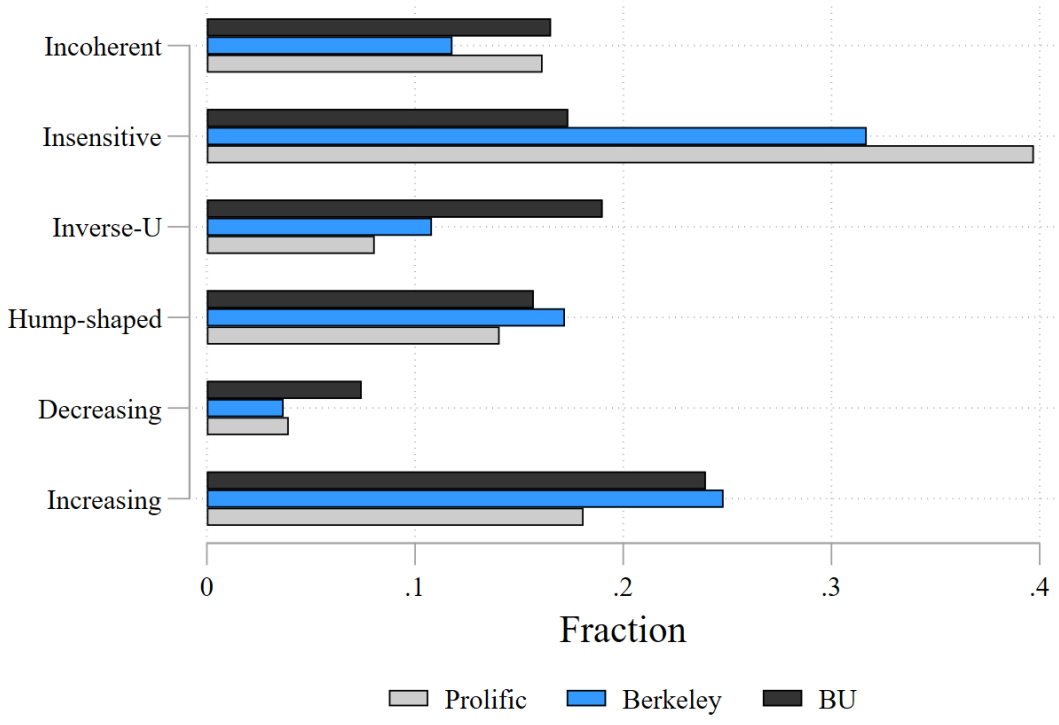
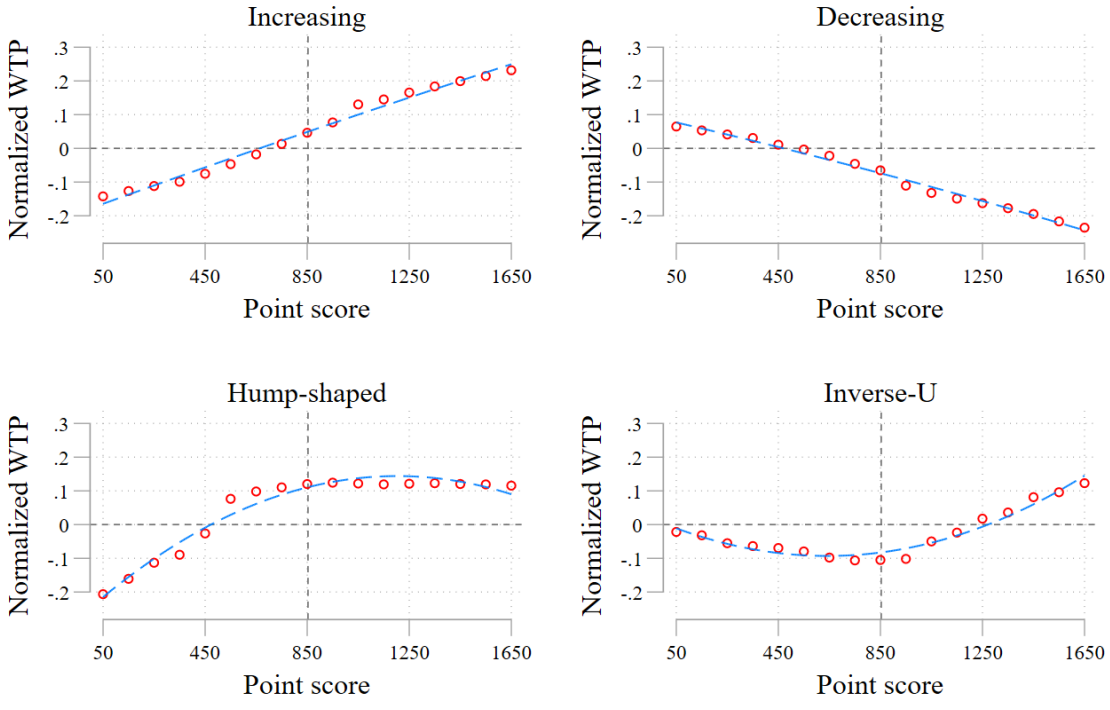


Figure 3 plots the average WTP statements by point score of the different non-trivial PRU types. The estimates suggest a correlation between the types and their moral standard. One interpretation is that inverse-U types have a kind of "winner takes all" preference for public recognition, where they only regard high point scores positively. Hump-shaped types on the other hand have more of a "last place aversion" preference that only regards very low scores negatively and stagnates as the score becomes above-average.

While BMM&T's method allows for the PRU identification on the individual level, the quality of any individual-level estimates will depend on how precisely participants answer the WTP questions. This was not the focus of BMM&T's original study and one may thus not want to rely too much on the individual-level estimates when interpreting BMM&T's results. However, the heterogeneity analysis results suggest that (i) there might be systematic "types" of individuals with different PRUs, (ii) these types can—in principle—be identified using BMM&T's technique, (iii) future research could use this technique in experiments more tailored at individual-level measurements to estimate the population type distribution and investigate implications for policy interventions and welfare.

Figure 3: Binned scatter plot of the average PRU, by PRU type



Note: The normalized WTP is a variable that takes on the value of 1 if the participant reported the maximum possible WTP for a given interval, a value of -1 of the participant reported the minimum possible WTP for a given interval, and a value of zero if the participant reported a WZP of zero. The maximum/minimum WTP was 25/-25 in the Berkeley and BU samples and 10/-10 in the Prolific sample. The vertical line denotes the average point score in the anonymous/charity only treatment and the x-axis labels display interval midpoints. The blue lines display the quadratic fit.

4 *Conclusion*

In this report, I replicate results from Butera et al. (2022). I can computationally reproduce results reported in the original paper. The results are further robust to different sample selection criteria, adding and removing controls, and correcting for coding errors. The results from the additional heterogeneity analysis complement those reported in the paper and could be further explored in future research.

A Adjustments made to the calculations originally reported in Table 5 of the paper

The original code included in the file `Build/Label/Rounds.do` assigns an order value between 1 - 3 to each observation in the sample, depending on how a participant went through the three treatments. For example, the file includes commands like `replace order=1 if round == "anom" & condition=="ARE"`, assigning an order value of "1" if a participant first worked on the "anom" task (with the other tasks being "recog" and "earn"). This file sometimes mistakenly assigns the wrong order to certain observations such as in the following extract taken from the code:

```
replace order=2 if round=="recog"&condition=="AER"
replace order=3 if round=="earn"&condition=="AER"
```

While a participant in condition AER first took part in "anom", then in "earn" and finally in "recog", the code assigns an order value of 2 "recog" and 3 to "earn".

I re-ran a corrected version of the code by replacing the order variable file with the following file, which should correctly assign order:

```
gen order = .
forvalues p = 1 / 3 {
  replace order = `p' if round == "anom" & substr(condition,`p',1)
     =="A"
  replace order = `p' if round == "recog" & substr(condition,`p',1)
     =="R"
  replace order = `p' if round == "earn" & substr(condition,`p',1)
     =="E"
}
```

B Adjustments made to the calculations originally reported in Table 6 of the paper

The Table 6 results are generated in the file `Analysis/Code/charity_reduced_form_analysis.do` provided in the replication package. The function used to calculate the term $-R''/R'(\bar{a}_{pop})\times SD$ looks as follows:

```
   capture program drop get_ci
   program define get_ci, rclass

     nlcom
        -2*_b[c.interval#c.interval]/(_b[interval]+2*_b[c.interval#c.interval]*`2'
        // measure of curvature

     matrix tempb = r(b)
     matrix tempv = r(V)
     loc x : di %4.3f tempb[1,1]
     loc lb : di %4.3f tempb[1,1] -
        (invnormal(.975)*(sqrt(tempv[1,1])))
     loc ub : di %4.3f tempb[1,1] +
```

```
      (invnormal(.975)*(sqrt(tempv[1,1])))

   return loc alpha "`x'"
   return loc ci : di "[`lb', `ub']"

   * Multiply estimate by SD of points in the anonymous round
   nlcom -2*`1'*_b[c.interval#c.interval]/_b[interval]

   matrix tempb = r(b)
   matrix tempv = r(V)
   loc x : di %4.3f tempb[1,1]
   loc lb : di %4.3f tempb[1,1] -
      (invnormal(.975)*(sqrt(tempv[1,1])))
   loc ub : di %4.3f tempb[1,1] +
      (invnormal(.975)*(sqrt(tempv[1,1])))

   return loc alpha2 "`x'"
   return loc ci2 : di "[`lb', `ub']"

end
```

This function takes two arguments, SD and $\bar{a}_{pop}$, and returns four arguments: $-R''/R'(\bar{a}_{pop})$, the 95%-CI of $-R''/R'(\bar{a}_{pop})$, $-R''/R'(\bar{a}_{pop}) \times SD$, and the 95%-CI of $-R''/R'(\bar{a}_{pop}) \times SD$. Since R is equal to

$$R(\bar{a}_{pop}) = CONSTANT + \beta[interval]\bar{a}_{pop} + \beta[c.intervalc.interval]\bar{a}^2_{pop},$$

we have

$$\begin{aligned} R'(\bar{a}_{pop}) =& \beta[interval] + 2\beta[c.intervalc.interval]\bar{a}_{pop}, \\ R''(\bar{a}_{pop}) =& 2\beta[c.intervalc.interval]. \end{aligned}$$

Therefore, the following line in the code above:

```
nlcom -2*`1'*_b[c.interval#c.interval]/_b[interval]
```

should be replaced by:

```
nlcom
   -2*`1'_b[c.interval#c.interval]/(_b[interval]+2*_b[c.interval#c.interval]*`2')
```

C Additional tables and figures

Table 4: Replication of Table 7 with corrected order variable

Panel A. Action-signaling model parameter estimates				
Sample	$\hat{\gamma}_1^a$	$\hat{\gamma}_2^a$	$\hat{\rho}^a$	$\hat{c}$
Prolific	0.12	-0.004	0.58	0.08
	[0.09,0.14]	[-0.005,-0.002]	[0.40,0.80]	[0.06,0.11]
Berkeley	0.30	-0.004	0.87	0.21
	[0.22,0.38]	[-0.011,0.003]	[0.63,1.15]	[0.14,0.34]
BU	0.38	0.002	1.59	0.30
	[0.19,0.53]	[-0.009,0.013]	[1.13,2.27]	[0.16,1.09]
Panel B. Characteristics-signaling model parameter estimates				
Sample	$\hat{\gamma}_1^\theta$	$\hat{\gamma}_2^\theta$	$\hat{\rho}^\theta$	$\hat{c}$
Prolific	1.29	-0.450	0.49	0.08
	[0.97,1.65]	[-0.753,-0.239]	[0.25,0.76]	[0.06,0.11]
Berkeley	1.35	-0.082	0.85	0.21
	[0.89,1.81]	[-0.324,0.048]	[0.55,1.17]	[0.14,0.34]
BU	1.26	0.026	1.66	0.30
	[0.26,2.57]	[-0.119,0.284]	[1.15,2.36]	[0.16,1.09]
Panel C. Predicted and actual effects of financial incentives (on attendance or points (00s))				
Sample	Model prediction		Actual	Pred. - Act.
Prolific	2.39		1.82	0.57
	[1.79,3.09]		[1.57,2.07]	[0.03,1.23]
Berkeley	2.34		1.77	0.57
	[1.47,3.53]		[1.34,2.24]	[-0.18,1.65]
BU	1.65		1.32	0.33
	[0.42,3.08]		[0.64,2.07]	[-0.79,1.66]

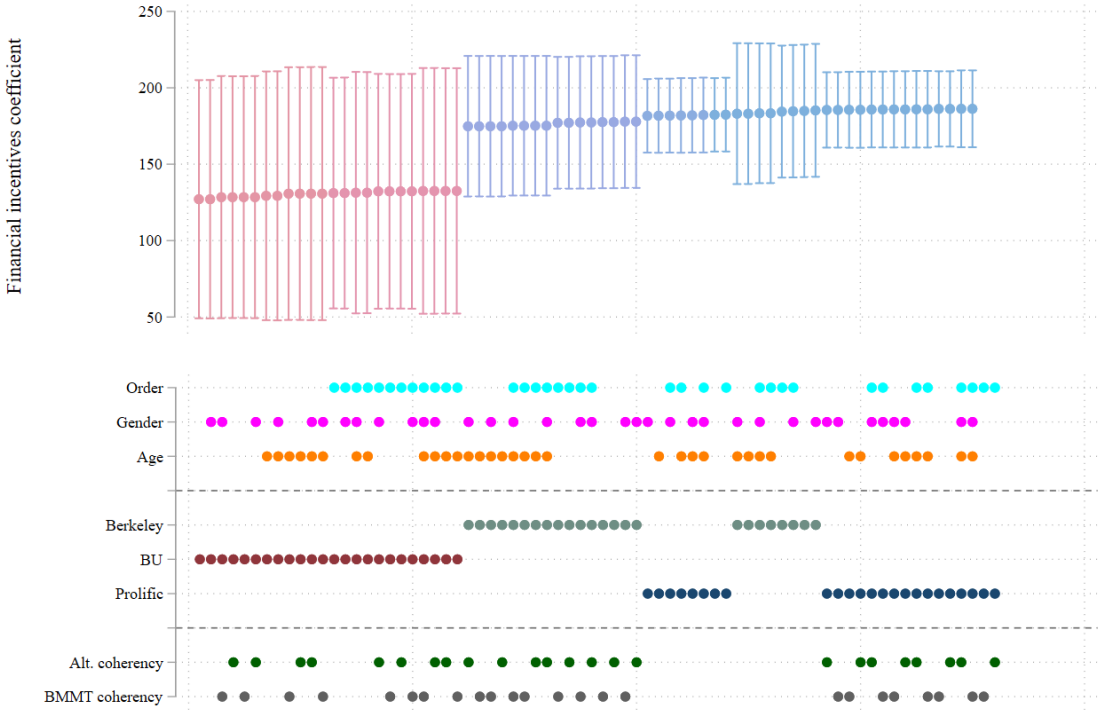
Note: The table omits results on the YMCA sample since the were not the focus of the replication.

Table 5: Replication of Table 9 with corrected order variable

Panel A. Action signaling model						
Row	Scenario	Parameter estimates			Image Payoffs	Change in attendance
		γ_1^a	γ_2^a	ρ^a	(1)	(2)
1.	Baseline (YMCA)	0.64	-0.020	1.85	-3.41	55.77%
2.	ρ from Prolific sample	0.64	-0.020	0.58	0.70	39.31%
3.	ρ from Berkeley sample	0.64	-0.020	0.87	-0.04	42.73%
4.	ρ from BU sample	0.64	-0.020	1.59	-2.38	52.12%
5.	Curvature from Prolific sample	0.64	-0.022	1.85	-3.51	57.06%
6.	Curvature from Berkeley sample	0.64	-0.010	1.85	-2.92	49.53%
7.	Curvature from BU sample	0.64	0.005	1.85	-2.25	41.93%
8.	ρ and curvature from Prolific sample	0.64	-0.022	0.58	0.66	38.82%
9.	ρ and curvature from Berkeley sample	0.64	-0.010	0.87	0.16	43.57%
10.	ρ and curvature from BU sample	0.64	0.005	1.59	-1.56	42.65%

Panel B. Characteristics-signaling model						
Row	Scenario	Parameter estimates			Image Payoffs	Change in attendance
		γ_1^θ	γ_2^θ	ρ^θ	(1)	(2)
1.	Baseline (YMCA)	1.28	-0.079	1.40	-1.18	47.55%
2.	ρ from Prolific sample	1.28	-0.079	0.49	0.50	40.25%
3.	ρ from Berkeley sample	1.28	-0.079	0.85	-0.12	43.11%
4.	ρ from BU sample	1.28	-0.079	1.66	-1.71	49.63%
5.	Curvature from Prolific sample	1.28	-0.078	1.40	-1.17	47.39%
6.	Curvature from Berkeley sample	1.28	-0.059	1.40	-1.07	45.64%
7.	Curvature from BU sample	1.28	0.019	1.40	-0.65	39.38%
8.	ρ and curvature from Prolific sample	1.28	-0.078	0.49	0.51	40.25%
9.	ρ and curvature from Berkeley sample	1.28	-0.059	0.85	-0.02	42.42%
10.	ρ and curvature from BU sample	1.28	0.019	1.66	-1.12	38.93%

Figure 4: Specification curve plot of the financial incentives coefficient



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, 1607–1636.

BÉNABOU, R. AND J. TIROLE (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

BURSZTYN, L. AND R. JENSEN (2017): "Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure," *Annual Review of Economics*, 9, 131–153.

BUTERA, L., R. METCALFE, W. MORRISON, AND D. TAUBINSKY (2022): "Measuring the Welfare Effects of Shame and Pride," *American Economic Review*, 112, 122–168.