SUPPLEMENTAL APPENDICES

Appendix A: Multicollinearity

We also assessed whether our regression results are subject to multicollinearity issues. As one might expect, the interaction between the *Treatment Count* and *Cash Offer* exhibits a relatively high correlation with the constituent terms, and produces a relatively high variance inflation factor (VIF) in turn. However, it should be kept in mind that high VIFs are not typically problematic when they result from correlations between interaction terms and their constituent variables. To demonstrate this in our setting, we apply the residual-centering approach of (**?**). The resulting regression yields very similar results to the baseline model, and the VIF values (reported along side the centered model values) are well within normal thresholds (see Table A.1). Ultimately, we conclude that collinearity is not influencing the results.

Table A.1 Treatment Count Model (LPM; DV = Convert)

Variable	Original Model	Residual Centering	VIF	1/VIF
1 *Treatment*	0.076** (0.032)	.0756** (.0317)	2.46	0.407
2 *Treatments*	0.060** (0.030)	.0603** (.0297)	2.46	0.4062
3 *Treatments*	0.109** (0.049)	.1213** (.0508)	1.85	0.5408
1 $Treatment \cdot Cash\,Offer$	0.052 (0.064)	.0115 (.0162)	3.42	0.2926
2 $Treatments \cdot Cash\,Offer$	0.086 (0.069)	.0216 (.0175)	3.43	0.2916
3 $Treatments \cdot Cash\,Offer$	0.211** (0.087)	.0509** (.0121)	2.46	0.4072
$Cash\,Offer$	−0.058 (0.057)	.0062 (.0057)	1.01	0.9936
Intercept	0.017 (0.017)	.0174 (.0169)	Mean VIF	2.44
Observations	323	323		
R^2	0.037	0.036		
F	3.85*** (7, 316)	3.82*** (7, 315)		

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Appendix B: Estimator Choice

One possible concern with our results is that they are somehow dependent upon bias or inconsistency of the LPM (**?**). It is important to note, however, first, that the typical concerns with bias and inconsistency of OLS and binary outcomes are not applicable to experimental treatment impact evaluations (**?**). Second, even in observational data, **?** have shown that the bias underlying LPMs is unlikely to be severe when the vast majority of predicted values a resulting model yields fall entirely within the 0-1 range. In the event that any predicted values do lie outside the feasible range, those authors propose the application of a trimming estimator. This estimator is a standard LPM that simply omits those observations holding infeasible predicted values. Employing this procedure notably only results in our excluding 5 observations from the original sample and, as can be seen in Table B.1, our coefficients remain essentially unchanged.

Table B.1 Trimmed OLS (LPM; DV = Convert)

Coefficient	Model (1)
1 $Treatment$	0.0870^{***} (0.0279)
2 $Treatments$	0.0711^{***} (0.0259)
3 $Treatments$	0.1171^{**} (0.0461)
1 $Treatment \cdot Cash\,Offer$	0.01938 (0.0217)
2 $Treatments \cdot Cash\,Offer$	0.0295 (0.0240)
3 $Treatments \cdot Cash\,Offer$	0.0295^{**} (0.0230)
$Cash\,Offer$	-0.0223 (0.0217)
Intercept	0.0054 (0.0075)
$Observations$	318
R^2	0.035
F	3.82^{***} $(7, 310)$

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Although a Logistic regression is often viewed as preferable when dealing with binary outcomes, because it has the desirable property of constraining predicted values to lie within the 0-1 interval, it is important to keep in mind that this estimator also has the *undesirable* property of yielding coefficients that are difficult to understand or interpret. This is true for two reasons. First, logistic regression deals with odds ratios, which often lack straightforward intuition, given their multiplicative nature. Second, the coefficients and standard

errors associated with interaction terms in Logistic Regression are not directly interpretable (**?**). That said, we also estimated a logistic regression model, the results of which are presented below in Table B.2. As can be seen, these results are qualitatively similar to those reported elsewhere, in terms of sign and statistical significance of the estimated coefficients.

Table B.2 Logit (DV = Convert)

Coefficient	Model (1)
1 $Treatment$	3.879^{***} (1.083)
2 $Treatments$	3.641^{***} (1.099)
3 $Treatments$	3.813^{***} (1.168)
1 $Treatment \cdot Cash\,Offer$	1.127^{***} (0.337)
2 $Treatments \cdot Cash\,Offer$	1.268^{***} (0.361)
3 $Treatments \cdot Cash\,Offer$	1.533^{***} (0.351)
$Cash\,Offer$	-1.161^{***} (0.324)
Intercept	-6.168 (1.036)
$Observations$	323
$Wald\ Chi^2$	29.14^{***}
$Pseudo\ R^2$	0.0659

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Appendix C: Randomization Checks

In this section, we report additional randomization checks, evaluating the orthogonality of cash offer and treatment assignment to one another, as well as between cash offer assignment and the clothing items that a subject brought to the buy-back procedure. Table C.1 shows the pairwise comparisons of the average cash offer assigned between alternative conditions, defined in terms of the number of treatments assigned. In Table C.2, we also report pairwise comparisons between each of the eight individual conditions, defined in terms of the unique combination of treatments assigned. All mean differences are statistically insignificant at the p<.05 level. These null results indicate that cash offer assignment was orthogonal to anthropomorphism treatment assignment.

Table C.1 Pairwise Comparisons cash offer and Number of Treatments

Test	Condition Comparison	*t*-stat	*p*-value
1	1 Treatment vs 2 Treatments	-0.228	0.820
2	1 Treatment vs 3 Treatments	-0.700	0.485
3	2 Treatments vs 3 Treatments	-0.527	0.600
4	1 Treatment vs 0 Treatments	-0.957	0.340
5	2 Treatments vs 0 Treatments	-1.117	0.266
6	3 Treatments vs 0 Treatments	-1.405	0.164

Finally, evaluating the correlation between the cash offer a subject was assigned and the number of clothes he or she was was seeking to sell (conditional on their progressing beyond the cash offer offer stage of the conversation), we again observed a statistically insignificant relationship ($p > 0.05$), implying that cash offer assignment was orthogonal to customer characteristics.

Table C.2 Pairwise Comparisons Cash Offer and Combination of Treatments

Test	Condition Comparison	*t*-stat	*p*-value
1	SP vs D	.093	0.93
2	SP vs H	0.172	0.86
3	SP vs SP & D	0.107	0.92
4	SP vs D & H	0.02	0.98
5	SP vs SP & H	-0.759	0.45
6	SP vs SP & D & H	0.554	0.58
7	SP vs Control	0.808	0.42
6	D vs H	0.052	0.96
7	D vs SP & D	0.00	1.00
8	D vs D & H	0.062	0.95
9	D vs SP & H	.771	0.44
10	D vs SP & D & H	-0.582	0.56
11	D vs Control	.093	0.54
12	H vs SP & D	-0.059	0.95
13	H vs D & H	-0.1370	0.89
14	H vs SP & H	-1.070	0.29
15	H vs SP & D & H	-0.777	0.44
16	H vs Control	0.445	0.77
17	SP & D vs D & H	-0.071	0.94
18	SP & D vs SP & H	-0.883	0.38
19	SP & D vs SP & D & H	-0.668	0.51
20	SP & D vs Control	0.710	0.48
21	D & H vs SP & H	-0.707	0.48
22	D & H vs SP & D & H	-0.505	0.61
23	D & H vs Control	0.717	0.48
24	SP & H vs SP & D & H	0.178	0.86
25	SP & H vs Control	1.717	0.09
26	SP & D & H vs Control	1.398	0.17

SP= Social Presence, D=Delay, H=Humor

Appendix D: Individual Treatments

The focus of the study is on the affects of anthropomorphism. However, from a practical standpoint, it is likely useful to also understand which of our treatments are most effective, individually. We therefore conducted additional analyses and another experiment on Amazon Mechanical Turk, aimed at addressing this question.

First, we report on our Turk experiment. In this experiment, we evaluated the desirability of individual treatment interventions in terms of subjects' reported perception of chatbot likeability. We limit this analysis to an Appendix, because it is not altogether clear whether responses from this artificial setting, in which subjects are paid to participate, would necessarily mirror results obtained in a field setting, wherein individuals organically opt into chatbot interactions. That said, results of this analysis may provide a useful indication of which anthropomorphic interventions may be particularly useful in practice.

We recruited 426 subjects on Mechanical Turk to interact with four versions of our chatbot, assigned at random: i. control, ii. social presence, iii. delay and iv. humor. We limited participation such that a given Turker could complete the HIT exactly once, to avoid concerns about interference and cross-over across conditions. After Turkers interacted with the their assigned chatbot, they were asked to rate the chatbot on the **?** enjoyable/unpleasant scale, which ranges from -100 to 100. We find that the humor treatment yields a significantly larger, positive effect than either the control or the delay treatment - Mann-Whitney U tests indicate statistical significance at conventional levels (p ¡ 0.05). We provide a visual depiction of the average likeability report by experimental condition in Figure D.1.

We also note here that the delay treatment yields significantly lower likeability than the control condition in this setting. As noted earlier, it is not clear whether this finding would also apply to our field setting. It should be kept in mind that crowd-workers are paid for their time. As such, our delay treatment in this setting not only manipulates anthropomorphism; it also implies that turkers are earning a lower effective wage. Moreover, we would note that we also do not account for interactions between different anthropomorphic treatments here. As such, it remains possible that delay can have a strictly positive, amplifying effect, as long as it is implemented in tandem with other anthropomorphic treatments.

Next, we revisited the data from our initial field experiment. A natural approach to consider is to simply remove our *Treatment Count* dummies and replace them with individual treatment dummies, along with all possible interactions. Unfortunately, such a model is under-powered, and yields a statistically insignificant

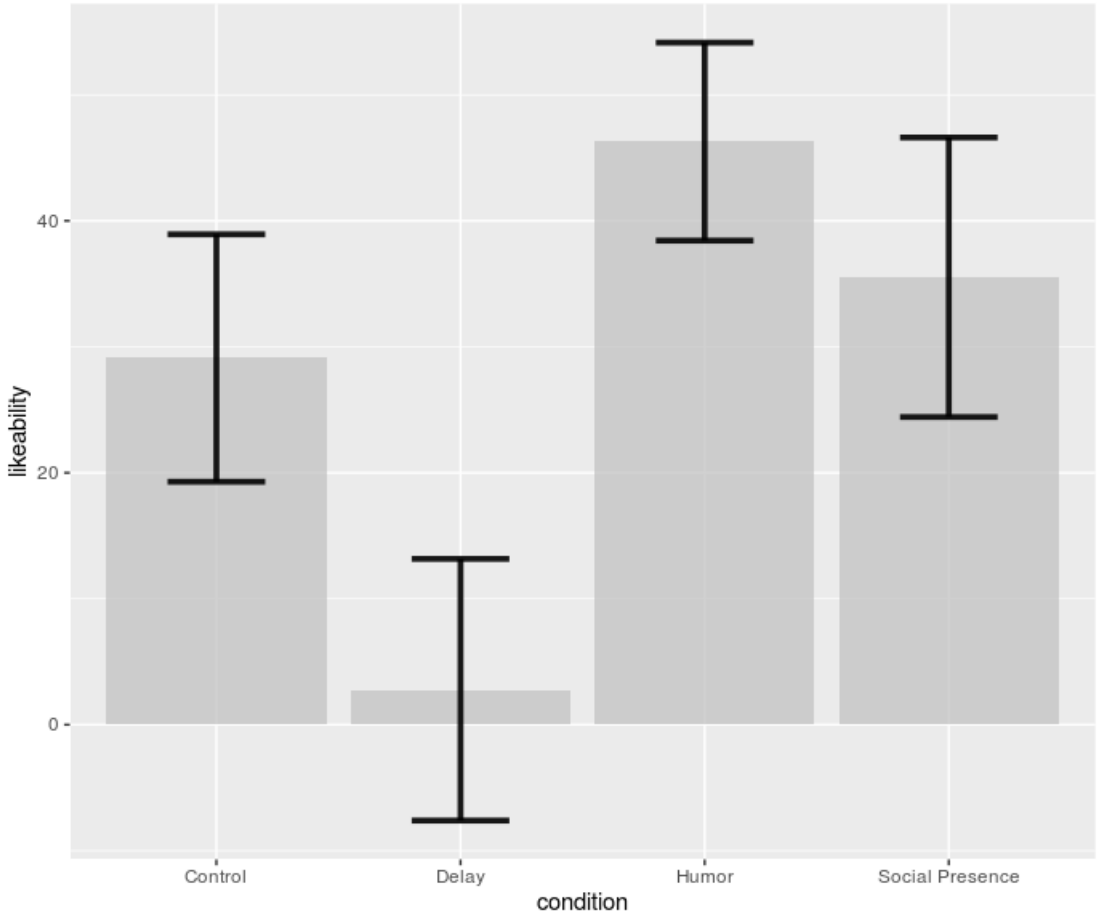

Figure D.1 Perceived Likeability - Individual Anthropomorphic Treatments

overall model fit. Accordingly, we considered a simpler regression specification, which ignores interactions between anthropomorphism treatments, and merely seeks to assess average main effects of each individual treatment, as well as their offer interactions. The model remains valid, of course, because all treatments and offer manipulations were varied exogenously.

Notably, this new model, estimated using the **?** trimming estimator, is statistically significant overall. The models yields an F-stat of 2.70 (7, 301), implying a p-value of 0.01 for overall model fit. The model results appear below in Table D.1. We observe results consistent with those seen in our Amazon Mechanical Turk study, above. That is, *Humor* has a significant, positive effect on conversion, whereas the coefficients on our two other interventions are null. Further, the effect of *Humor* is significantly larger than the effect of *Delay* ($p = 0.07$). Additionally, we see that *Delay* has a statistically significant interaction with *Cash Offer*,

suggesting it has a particular influence on offer sensitivity. Of course, these results are far from conclusive; additional work should be pursued to identify the ideal anthropomorphic interventions for retail settings.

Table D.1 LPM (DV = Convert)

Coefficient	Trimmed LPM (1)
$Delay$	-0.012 (0.030)
$Humor$	0.067^{**} (0.032)
$SocialPresence$	0.018 (0.031)
$Delay \cdot Cash\,Offer$	0.156^{**} (0.054)
$Humor \cdot Cash\,Offer$	-0.002 (0.051)
$SocialPresence \cdot Cash\,Offer$	0.091 (0.053)
$Cash\,Offer$	-0.103 (0.057)
$Observations$	309
$F-stat$	$2.70^{*}(7, 301)$
R^2	0.052

Note: Robust SEs; ** $p<0.05$, * $p<0.10$.