Web Appendix **Not to be Published**

1. Belief-dependent lie aversion in our decision problem: Predictions for a particular utility function

Suppose that utility can be described by function $U = x_i(z) - \gamma \max[0, I(z) - \mu(z)]$, where all variables are defined as in Section 2, and γ is an idiosyncratic sensitivity parameter randomly distributed over some non-negative interval according to a differentiable cumulative density function *F*. We posit that the sender maximizes utility given beliefs and uses Bayes' rule, where possible, to define the second-order beliefs. Predictions in our decision problem are as follows. First, never lie when the circle is green. Second, since the monetary gain from lying is 1 in our decision problem, it follows that the sender will lie on seeing the blue circle if $\gamma \cdot (1 - \mu^G) < 1$; note that the fraction of senders $f(\mu^G)$ for which this condition is satisfied coincides with $F(1/(1 - \mu^G))$, taking value 1 if $\mu^G = 1$. If the support of *F* includes 0, Bayes' rule will always be defined. A sufficient condition for equation (2) to have a unique solution is that function $h(\mu)$ from footnote 11 is strictly increasing so that

$$f'(\mu^G) < \frac{1 - p_B + f(\mu^G) p_B}{(1 - \mu^G) \cdot p_B}$$

This expression is not tremendously intuitive, but does make precise the idea that the distribution of sensitivity (and therefore of threshold gains for honesty) must not be "too concentrated". It is satisfied by many common distributions. For instance, if γ distributes uniformly on $[0, \Gamma]$ then $F(\gamma) = \gamma / \Gamma$, implying that $f(\mu^G) = 1/(\Gamma(1 - \mu^G))$, so that equation (1) in Section 3 of the paper implies $f = (1 - p_B)/(\Gamma(1 - p_B) - p_B)$ – this may be a corner solution if Γ is not sufficiently large. A computational analysis shows similar results for the exponential, logistic, or normal distributions. For instance, some examples of the fixed point for different parameterizations of a normal distribution are shown below.

Normal dist. Parameters		Equilibrium *f* (predicted lie rate)	
Mean	Standard Dev	$p_{\text{B}} = 0.25$	$p_{\text{B}} = 0.75$
0.5	1	0.775971	0.999767
1	1	0.576148	0.998632
1.5	1	0.35084	0.993437
2	1	0.173012	0.972392
2.5	1	0.069877	0.859787
3	1	0.02317	0.027608
3.5	1	0.006246	0.006563
4	1	0.001352	0.001368

2. A control treatment with the specific response method: Procedures and data

We report here data from a control treatment that used the specific response method in order to show that our main results are not an artifact of the use of the strategy method. To motivate this point, recall that subjects in the High and Low treatments had to indicate the message to be sent in any possible contingency; i.e., they made hypothetical decisions for both circle colors before discovering the true color of the circle. Given the hypothetical nature of these decisions, one might argue that emotions are less vivid in this case, and that could have an effect on behavior. To check for this, we ran a control treatment without the strategy method, which coincided with our High treatment in everything except that subjects made their choice after seeing the randomly selected circle color in their screens, and that beliefs about deception were conditional on having seen the blue signal. We focused on the High treatment simply because lies are most likely when the signal is blue; recall also that previously we found no significant differences in honesty across treatments.

A total of 40 subjects participated in this control treatment, and the distribution of gender and major was similar to our two other treatments. In effect, a Mann-Whitney test of the hypothesis of equal gender distributions yields a p-value of 0.700, while Fisher's exact test of the equality of the major distributions yields a p-value of 0.796. Further, our main results are replicated in this control treatment. First, among the 30 subjects who saw the blue circle in their screens, 40 percent sent the truthful message 'blue'. This is not significantly different than the analogous rate in the High treatment (Mann-Whitney p = 0.909). We also observe a correlation between honest behavior and beliefs. Subjects who sent a false 'green' message reported both first and second-order beliefs of deception that were significantly higher than those reported by subjects who chose to send a truthful ‘blue’ message (first order: $p < 0.001$; second order: $p < 0.005$). Table W1 shows subjects’ average beliefs about deception depending on history of play (i.e. the color of the circle observed, and the message sent afterwards). First and second-order expectations were again strongly correlated ($r = 0.786$, $p < 0.0001$).

Circle color	Average beliefs	Message sent		Mann-Whitney p
		Green	Blue	
Blue	first-order	90.11	57.92	0.0008
	second-order	89.78	63.08	0.0010
	N	18	12	
Green	first-order	80.75	52.50	0.087
	second-order	81.38	67.50	0.290
	N	8	2	
Note: Mann-Whitney test of equality of beliefs across messages sent.				

Table W1: Average beliefs about the frequency of dishonesty, conditional on history of play