Games and Economic Behavior 102 (2017) 445–454

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

A must lie situation – avoiding giving negative feedback ☆

Uri Gneezy [a,b], Christina Gravert [c], Silvia Saccardo [d,*], Franziska Tausch [e]

[a] Rady School of Management, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[b] Center for Research in Experimental Economics and Political Decision-Making, University of Amsterdam, 1018 WB, Amsterdam, The Netherlands
[c] Department of Economics, University of Gothenburg, Vasagatan 1, SE 411 24 Göteborg, Sweden
[d] Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
[e] Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany

ARTICLE INFO

Article history:
Received 11 June 2016
Available online 30 January 2017

JEL classification:
D03
C91
D83

Keywords:
Lying
Feedback
Overconfidence
Updating
Laboratory experiment

ABSTRACT

We examine under what conditions people provide accurate feedback to others. We use feedback regarding attractiveness, a trait people care about, and for which objective information is hard to obtain. Our results show that people avoid giving accurate face-to-face feedback to less attractive individuals, even if lying in this context comes at a monetary cost to both the person who gives the feedback and the receiver. A substantial increase of these costs does not increase the accuracy of feedback. However, when feedback is provided anonymously, the aversion to giving negative feedback is reduced.

"I don't want any yes-men around me. I want everyone to tell me the truth – even if it costs him his job."
[Samuel Goldwyn]

1. Introduction

Feedback is crucial to learning. The transmission of information from an informed agent to a receiver who might benefit from it is studied theoretically in the standard principal-agent model in economics (Crawford and Sobel, 1982; Prendergast, 1993; Levitt and Snyder, 1997; Morris, 2001; Olszewski, 2004; Ottaviani and Sørensen, 2005a, 2005b). The question in this literature is how to design incentives contracts such that the agent with the private information will send an honest signal to the principal. We expand the discussion by studying cases in which the agent might suffer psychological costs from sending a negative signal, and hence avoids it.

Whereas a growing literature suggests that people have costs associated with lying (see for example Gneezy, 2005; Charness and Dufwenberg, 2006; Mazar et al., 2008; Dreber and Johannesson, 2008; Sutter, 2009; Erat and Gneezy, 2012; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2016), in some settings individuals may have costs associated with telling

☆ This research was conducted under IRB# 130855 at University of California San Diego.

* Corresponding author.
E-mail address: ssaccard@andrew.cmu.edu (S. Saccardo).

http://dx.doi.org/10.1016/j.geb.2017.01.008
0899-8256/

the truth, even if being honest is payoff maximizing for both the principal and the agent. These costs could be the time and effort required to give accurate feedback, or, as we suggest in this paper, the psychological costs that arise from delivering negative messages. This last effect is what we call a "must lie situation."[1] The aversion to telling the truth that is perceived as negative for the receiver could thus offset the costs associated with lying that have been documented in other studies.

In many cases, feedback can help people achieve better outcomes. Consider the thousands of individuals who show up for singing or comedy castings, or who dedicate their life to research or business ideas. They mostly need to rely on their own, likely biased perception of themselves and possibly on too flattering feedback from family and friends. Receiving honest feedback from other individuals could prevent them from wasting time and resources and save them from frustration. Similarly, consider a person on the job market who keeps applying to jobs that he might be qualified for on paper, but is considered unsuitable based on less tangible character traits displayed during the interview process. He might talk, act or dress in a way that displays a low work ethic or just does not fit with the company image. Honest feedback about his personal characteristics could help such a person to revise his application strategy and thus to be more successful in the process – either by applying to companies that are a better fit to his personality or by adapting his behavior to the companies he is applying for. Consequently, a lack of feedback could lead to frustration, extended unemployment spells and superfluous investment into further education.

In order to study the provision of feedback, we designed a novel experiment in which participants are asked to give feedback to others on their level of attractiveness. We decided to use attractiveness as a proxy for similar less-tangible traits that could be subject to feedback for four main reasons. First, whereas for some attributes people have a good knowledge of their relative rank (e.g., height), the feedback regarding own attractiveness is noisy and relies on indirect measures such as success in dating. Hence, receiving accurate feedback could be very informative. Second, attractiveness is an attribute most people care about a lot, and thus receiving an informative negative signal could hurt. Third, attractiveness is correlated with economic success (Solnick and Schweitzer, 1999; Möbius and Rosenblat, 2006). Fourth, attractiveness can be judged within seconds in an experimental setting, while other traits might only be revealed after an extensive interaction.

In our experiment we asked groups of men and women to rank the members of the opposite sex by attractiveness. We then incentivized participants to correctly judge the rank of another participant of their own sex in private, and compared these assessments to those provided in a treatment in which the attractiveness judgments were provided to the assessed individual as face-to-face feedback. We find that participants are reluctant to provide honest negative face-to-face feedback to other people even if lying in this context comes at a cost to both the feedback provider and the receiver. Further, we find that a considerable increase in these costs does not change the accuracy of individuals' feedback provision.

One reason for the avoidance to provide negative face-to-face feedback could be a personal cost. The receiver of the information might decide to "shoot the messenger" – blame the (innocent) carrier of bad news – which could result in monetary or image costs for the sender (Akerlof and Kranton, 2000; Ariely et al., 2009). Psychologists, starting with Freud (1991), studied this phenomenon, arguing that people may blame the messenger for the message as a mechanism to fight feelings of powerlessness and a lack of control. Alternatively, the avoidance of giving negative feedback could rise from trying to shield the receiver from negative information that could hurt. That is, individuals may experience negative utility from providing signals that they anticipate to be painful to the receiver.

To better understand what drives the reluctance to provide honest negative feedback that we find in our experiment, we ran a treatment in which feedback provision is anonymous, i.e., the identity of the feedback provider is not revealed to the receiver. We find that participants provide more honest feedback when their anonymity is guaranteed as compared to when their identity is revealed. This suggests that the reluctance to give face-to-face feedback to less attractive individuals is driven by unwillingness to be identified as the messenger of bad news.

Taken together, our results suggest that the costs of being identified as the messenger of negative feedback offset the potential costs of lying that have been identified in previous research. In order to protect their social image individuals are willing to tell a payoff minimizing lie, even at high cost.

2. Experimental design

2.1. The setting

The experiment consists of four treatments: Judgment, Face-to-Face (F2F) Feedback, F2F High Stakes Feedback, and Anonymous Feedback. Participants took part in four stages. The first two stages are procedurally the same in all treatments, while the experimental design differs in stages three and four. A treatment overview is depicted in Table 1.

In order to balance between having enough power and being able to recruit enough participants for all our treatments, we decided to recruit 100 participants (50 men and 50 women) per treatment. Each session consisted of 20 participants, 10 men and 10 women. Upon arrival to the laboratory, men were instructed to line up on one side of the room while women formed a line on the opposite side, such that the two groups faced each other. Then participants received the instructions as

[1] A funny anecdote from "Seinfeld" inspired the name of our paper: Jerry and Elaine are invited to see their friends' baby. One look and they both agree the baby is "the ugliest baby you have ever seen." They of course do not tell this to the proud parents. Jerry's insight later is "And, you know, the thing is, they're never gonna know, no one's ever gonna tell them. . .it's a must lie situation."

Table 1
Treatments.

Stage\Treatment	Judgment	F2F	F2F High Stakes	Anonymous
1 Ranking opposite group	✓	✓	✓	✓
2 Guessing own rank	✓	✓	✓	✓
3 Guessing other's rank	Private judgment	F2F Feedback	F2F Feedback	Anonymous Feedback
4 Updating	Ranking own group	✓	✓	✓

well as an ID sticker, which they were asked to wear visibly on their chest. Women received ID letters from FA to FJ; men received ID letters from MA to MJ. After putting on the stickers, participants were asked to start reading the instructions for stage one (all the instructions are reported in Online Appendix D).

In the first stage of the experiment participants were asked to rank the members of the opposite group by attractiveness from 1 (the most attractive person) to 10 (the least attractive person), such that each person in the other group received a different rank. The ranks given to each participant in a group were added up and the resulting sums were ordered from the lowest to the highest value. Based on this order, we created an aggregate attractiveness ranking for the group from 1 to 10, such that each participant received a different rank. In the unlikely case of a tie, ranks were determined by the flip of a coin. To incentivize accuracy in the ranking, we promised participants a monetary reward if their rankings matched the aggregate ranking for at least five people. This procedure allowed us to incentivize decisions and provided us with a ranking that mainly mirrors generally perceived attractiveness instead of individual-specific, taste-based assessments. This reward was specified as $10 in the Judgment, F2F and the Anonymous treatment, and as $50 in the F2F High Stakes treatment.

After everyone had completed the first task (and before knowing whether their own ranking was in line with the group's ranking), we instructed participants to continue with the second stage. In this stage we asked participants to form a circle with their own sex group, so that all other group members were clearly visible to everyone. We reminded them of the aggregate attractiveness ranking provided by participants of the opposite sex group, and asked participants to guess their own rank in the aggregate ranking. That is, we asked participants to guess how the other group had ranked them. Participants received $10 (or $50 in the F2F High Stakes treatment) if their guess matched their actual position in the aggregate ranking. Without knowing whether their guess of their own rank was correct, participants were then asked to turn to the next page of their instructions to continue with stage three.

In stage three participants guessed the rank in the aggregate ranking of another same sex participant in their group. In the Judgment treatment we told participants that their guess would not be revealed to the participant whose rank they were assessing, and that they would be paid $10 for a correct guess. In the remaining treatments participants had to send a message to their counterpart with their assessment. Each participant gave feedback to one participant and received feedback from a different participant afterwards. In particular, participants were asked to write their guess in a message that was delivered to the receiver by the experimenter (see in the Appendix a sample of the message individuals had to complete). The message stated *"My guess about participant (1D)'s position in the aggregate ranking (1–10): __"*, and participants had to enter a number from 1 to 10.

In the F2F and the F2F High Stake treatment senders knew that in addition to the rank assessment they gave, the receiver would also know their identity; the sender's ID was pre-written on the message. In the Anonymous treatment information about the identity of the feedback provider was not given. To incentivize participants, we informed them that in stage four, after receiving the message, each participant would have the opportunity to update the guess of the own rank provided in stage two. If a participant guessed his/her rank correctly in stage four he/she would receive $10, and so would the person who provided feedback to him/her. In the F2F High Stakes treatment the incentives for both participants were increased to $50. After sending their messages, participants also received a similar message from another participant of the same sex with feedback on their ranking. Again, the ID of the participant who sent the message was visible on the message sheet only in the F2F treatments.

In stage four, upon receiving their message, participants in the F2F, F2F High Stakes and Anonymous treatment decided whether to update their personal rank prediction from stage two. Participants were notified that a correct, unaltered guess in both stages two and four would only be rewarded once and that the guess provided in stage four overruled the one provided in stage two. If the participant did not guess correctly in stage four, he/she and the person who sent him/her feedback did not receive additional money in this stage. In the Judgment treatment, given that participants did not receive any signal about their attractiveness, we did not ask them to update their ranking. We instead asked them to provide a complete attractiveness ranking of their own group including themselves. The incentive structure was identical to that of stage one.

2.2. Procedural details

We conducted the experiment with undergraduate students in a large university. A total of 400 participants (50% female) participated in the experiment, with 100 participants in each of the 4 treatments. Our sample consists of 62% Asians, 14% Caucasians, 10% Hispanics, and 9% indicated a different ethnicity. For 6% of the sample we do not have any ethnicity information. The average age is 20.5 years with a standard deviation of 1.9.

Participants were recruited through the online participant database of the university. In order to ensure a total of 20 participants per session, we recruited 30 participants to the lab (half men and half women). We then selected 10 men and 10 women at random for the experiment. All the remaining participants were dismissed after receiving a $5 show up fee. No participant took part in more than one session. We ran the experiment using pen and paper in the spring and fall of 2013 and winter of 2014. On average, each session lasted around 20 minutes. The average payment for the experiment was $9.92 plus a $5 show-up fee. To guarantee confidentiality to all participants, individuals were communicated their total earnings without being told their earning for any given stage of the experiment. None of the participants succeeded in earning money for all stages of the experiment.

3. Results

In the following, we first present the results on participants' self-assessment. We then turn to their evaluation of others' attractiveness and the resulting updating behavior. Throughout the analyses we will use the aggregate ranking computed using participants' evaluations of all opposite sex participants during stage one as a measure of participants' actual rank.

To make sure that the aggregate ranking is a meaningful indicator of their relative attractiveness, we first verify that participants' evaluation of other peoples' generally perceived attractiveness is similar. We find that the rankings of participants of the opposite sex are highly correlated among group members in the $10 incentive treatments (Cronbach's Alpha for men $= 0.85$ and for women $= 0.88$, $N = 298$).[2,3] The results are similar in the F2F High Stakes treatment (Cronbach's Alpha for men $= 0.90$ and for women $= 0.92$, $N = 100$). Hence, even though taste might differ across individuals, we observe a high degree of agreement regarding the perceived relative attractiveness of others.

3.1. Self-assessment

Feedback is particularly useful in situations where one's own perception and the perception of others differ. A necessary prerequisite for our study is thus a discrepancy between individuals' self-assessment and their rank in the aggregate attractiveness ranking. Since the self-evaluation stage of the experiment (stage two) was identical across the Judgment, F2F and Anonymous treatment, we pool all the observations from these treatments for this analysis. We present the results of the F2F High Stakes treatment separately, as participants in this treatments faced different incentives for this stage.

As can be seen in the distribution of self-assessments depicted in Fig. 1A, there is a considerable drop in the fraction of participants who rank themselves as 7 or higher. While ranks 1 to 6 are guessed by a fraction of people that is equal or larger than the expected 10% per rank, the fraction of participants who indicate any of the ranks above 6 is significantly lower than the expected 10% per rank. Overall, only 7% of the participants guessed that they were ranked among the less-attractive in the group as opposed to the expected 40% (test for proportions, $z = -11.37$, $p < 0.001$). We will use 7 as a threshold for the distinction of "attractive" (ranks 1–6, black bars) and "less-attractive" ranks (ranks 7–10, gray bars) throughout our analyses.[4]

In order to detect any possible bias in individuals' self-assessment we compare the distribution of guesses of participants' own rank to the distribution we would expect under full information. If participants were perfectly aware of their position in the ranking, we should observe a discrete uniform distribution of guesses, because each rank can only be awarded once per group. However, as can be seen in Fig. 1A, the distribution of the self-assessments when individuals are incentivized with $10 is skewed to the left and significantly different from a uniform distribution ($\chi^2(9) = 95.20$, $p < 0.001$, $N = 298$).[5] Participants assign a mean rank of 3.85 to themselves, instead of the expected 5.50.

When investigating the extent to which individuals' perception of their own attractiveness deviates from the aggregate ranking provided by the other participants in stage one (i.e., their actual rank), we find that on average individuals rank themselves as 1.63 ranks more attractive than their actual rank (SD $= 2.35$, $N = 298$). A Wilcoxon signed rank (WS) test confirms that this difference is statistically significant ($z = -9.97$, $p < 0.001$). Overall, 63% of the participants guess that they are more attractive than others perceive them to be. Those individuals on average deviate by 3 ranks from their actual rank (SD $= 1.73$, $N = 189$). Hence, we observe a considerably biased self-assessment in our sample.

This overconfidence is well documented in the psychology and economics literature (e.g. Svenson, 1981; Englmaier, 2006; see Moore and Healy, 2008 for a review, or Balafoutas et al., 2012 and Burks et al., 2013 for applications). We also observe that this overly positive self-evaluation is true for both sexes: there are no differences in the self-assessment distributions of men and women ($\chi^2(9) = 12.33$, $p = 0.195$) or in the average deviation of their self-assessment from their actual rank (Mann Whitney (MW) test, $z = 1.19$, $p = 0.235$). Looking at the extent to which participants' self-assessments deviate from

[2] The number of observations is 298 instead of 300 as one participant indicated numbers instead of ID letters in this stage and a second participant left out more than one ID number in his ranking. Out of all 400 participants in all four treatments 18 left out one ID letter in their ranking and repeated another one instead. In this case we randomly assigned the missing ID number to one of the rank positions with the repeated ID.

[3] Cronbach's Alpha is a measure of internal consistency. It measures the average correlation of the individual rankings we use to create the aggregate ranking. The higher the score, the more reliable is the generated scale.

[4] In Online Appendix B we present robustness checks in which we (i) split the sample at the median using a threshold of 6, as well as (ii) split the sample at the 70th percentile using a threshold of 8. Splitting the sample at the alternative thresholds does not qualitatively change our main results.

[5] Two participants did not indicate their self-assessment, leaving us with a total of 298 participants for this analysis.

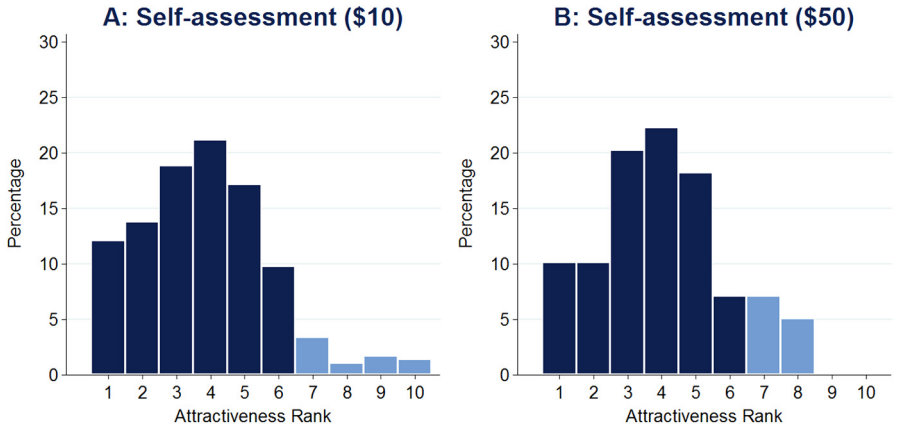

Fig. 1. Distribution of participants' guesses about their own attractiveness.

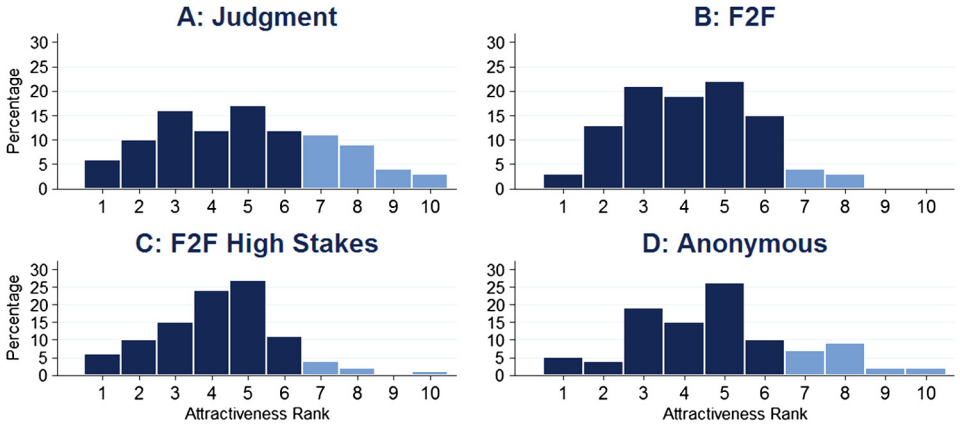

Fig. 2. Distribution of participants' guesses about others' attractiveness.

their actual rank, we see that on average the self-assessment of the less attractive individuals is more biased (average deviation = 3.58, SD = 1.90) than that of the attractive individuals (average deviation = 0.36, SD = 1.63). The difference is statistically significant (MW test, $z = 11.68$, $p < 0.001$). While the difference in deviations is considerable, we cannot exclude that a part of it is due to ceiling effects for the most attractive individuals.

The overconfidence in self-evaluation comes with a monetary cost to the participants. For a small amount of money, participants might get more benefit from deceiving themselves than from being honest about their attractiveness ranking and earning $10 for a correct guess. According to this explanation, when the cost of an overconfident answer goes up, the net benefit of self-deception will decrease and possibly become negative. This explanation predicts that the level of overconfidence would decrease as the cost of it increases. To test this hypothesis, we compare the results above with the results of the F2F High Stakes treatment, in which the cost of providing a biased self-evaluation is five times higher. The main result is that participants do not become more precise when the incentives to do so are increased (see Fig. 1B). Comparing the distribution of participants' guesses of their own rank observed in this treatment to the one observed in each of the $10 incentive treatments shows that increasing the incentive does not change the distribution of self-assessments ($\chi^2(9) = 10.12$; 8.02; 10.91, $p \geq 0.282$). While these incentives may yet be too low for overcoming overconfidence, a more plausible explanation is successful self-deception. Participants may honestly believe they are more attractive.

Result 1: Participants are overconfident in their self-evaluation irrespective of the size of their financial incentive.

3.2. Judgment and feedback

Only if individuals are able to assess the attractiveness rank of another same-sex person objectively when incentivized, feedback can be useful to correct the bias in individuals' self-assessment. We therefore first explore the accuracy of individuals' private judgments in stage three. In an objective judgment distribution, we should observe all ranks from 1 to 10 equally often. The distribution of guesses (see Fig. 2A) in the Judgment treatment is not significantly different from a discrete uniform distribution ($\chi^2(9) = 11.00$, $p = 0.275$, $N = 100$). The average guessed rank of another same-sex participant in this treatment is 4.9.

We further look at the deviation of individuals' assessment of the other's attractiveness. A negative deviation indicates that participants received positive feedback compared to their self-evaluation, whereas a positive deviation indicates that individuals received negative feedback compared to their self-evaluation. We find that on average individuals ranked their counterparts 0.56 ranks better than their actual rank (SD = 2.28, $N = 100$). This deviation is significantly smaller than the 1.68 deviation (SD = 2.07) when participants assess their own attractiveness (WS test, $z = -3.03$, $p = 0.001$). These results suggest that while participants tend to considerably overestimate their own attractiveness, they perform better when asked to judge others.[6]

When we analyze the behavior of participants matched with attractive (ranks 1–6) versus less attractive (rank 7–10) counterparts, we see that this deviation is not significantly different from zero for the former group (average deviation = −0.33, SD = 2.18, WS test, $z = -1.17$, $p = 0.243$), while it is significantly different for the latter group (average deviation = 1.90, SD = 1.71, WS test, $z = 4.95$, $p < 0.001$). While participants are on average precise when they evaluate attractive counterparts, they tend to evaluate less attractive counterparts as slightly better looking than they actually are. Nevertheless, when we compare how less attractive individuals assess themselves to how another same sex participant assesses them we find that in the latter case the deviation from the actual rank is significantly smaller (3.35 versus 1.9, MW test, $z = -3.58$, $p < 0.001$). Finally, we do not find evidence that the attractiveness of a participant affects the way he/she assesses an attractive/less attractive counterpart (MW tests, $z = 0.098$; 0.987, $p \geq 0.323$).

Next, we present the results of the F2F treatment. If feedback is honest, the messages sent to others in the third stage of the experiment should not differ from the evaluations provided in stage three of the Judgment treatment. This is not what we find in the F2F treatment. Participants send their counterparts overly positive messages (mean rank 4.2, $N = 100$; Fig. 2B). We find that the distributions of ranks in the Judgment and the F2F treatment are significantly different ($\chi^2(9) = 17.89$, $p = 0.036$). In line with our "must lie" hypothesis, this result is driven by the participants who are matched with a less attractive receiver. The distribution of ranks provided by these participants significantly differs between treatments ($\chi^2(8) = 18.99$, $p = 0.015$). By contrast, the distribution of ranks provided by participants matched with attractive individuals does not differ ($\chi^2(8) = 5.91$, $p = 0.657$). While 27% of the participants in the Judgment treatment evaluated their counterpart as less attractive, only 7% in the F2F treatment did so ($\chi^2(1) = 14.17$, $p < 0.001$). Strikingly, none of the participants of the F2F treatment gave ranks 9 or 10 as feedback. Thus, individuals appear to avoid evaluating others as less attractive when this information is delivered face to face to the other person.

We define feedback as dishonest if the deviation of individuals' assessment from the actual rank significantly differs from the deviation observed in the Judgment treatment. Limiting our analysis to participants who were matched with an attractive individual, in the F2F treatment this deviation is on average −0.15 (SD = 1.74), which does not significantly differ from the −0.33 (SD = 2.18) observed in the Judgment treatment (MW test, $z = -0.34$, $p = 0.737$). This suggests that participants give honest feedback to attractive counterparts. However, when matched with a less attractive participant, the average deviation is significantly larger in the F2F treatment than in the Judgment treatment (3.48 (SD = 1.72) vs. 1.90 (SD = 1.71) respectively (MW test, $z = -3.68$, $p < 0.001$).

The results discussed so far are confirmed by the OLS regression reported in Table 2 column (1), in which we regress the deviation of individuals' assessment from the actual rank on treatment dummies and a dummy for whether a participant's rank is defined as less attractive. Column (1) shows that the deviation of individuals' assessment from participant's actual rank is positive and significantly higher in the F2F treatment than in the Judgment treatment, suggesting that individuals in this treatment provided significantly more positive assessments. Further, as shown in the non-parametric tests, less attractive individuals are more likely to be assessed as more attractive than their actual attractiveness rank. In column (2), we interact the treatment dummies with the dummy variable indicating whether individuals are matched with less attractive receivers. This regression reveals that the treatment difference observed in column (1) is driven by individuals who are matched with a less attractive participant. The significant interaction between the F2F treatment and the attractiveness dummy suggests that the deviation of individuals' assessment from participant's actual rank is larger if the receiver is less attractive. That is, individuals inflate their feedback when matched with a less attractive receiver. This result is robust to including demographic controls (see column (3)). In Table A1 of Appendix A we additionally report the results of an OLS regression on the deviation of the assessment without imposing an attractiveness threshold, and find consistent results. See Appendix A for additional analyses.

Although the senders might intend to "be nice" by giving more flattering feedback, such kindness comes at a monetary cost to themselves and to their counterpart. As with self-confidence, while people might be willing to provide overly nice feedback when the price for doing so is $10, they might be less willing to do so when the price is $50. This is not the case in our experiment. By looking at the result of the F2F High Stakes treatment, we observe that participants do not send more honest messages (mean rank 4.22, $N = 100$). Only 7% of the participants in this treatment told their counterpart that they are less attractive (Fig. 2C). The distribution of feedback in this treatment does not differ from the distribution in the F2F treatment with low stakes ($\chi^2(8) = 5.30$, $p = 0.725$), while it significantly differs from the distribution of ranks observed in the Judgment treatment ($\chi^2(9) = 19.07$, $p = 0.025$). Again, the difference in the distribution of ranks is driven by the

[6] Men and women appear to have a similar perception of attractiveness. Using the data collected in the last stage of the Judgment treatment, we pool the rankings assigned by individuals' own group with the rankings assigned by the opposite sex group. Merging men's and women's opinions about each individual provides a Cronbach's alpha of 0.94 for the men's ranking and a Cronbach's alpha of 0.95 for the women's ranking.

Table 2
Deviation of the assessment of others' attractiveness rank from their actual rank by treatment.

Deviation	(1)	(2)	(3)
F2F	0.74***	0.18	.19
	(0.24)	(0.27)	(.33)
F2F High Stakes	0.72***	0.10	.17
	(0.21)	(0.29)	(.31)
Anonymous	0.10	−0.28	−.20
	(0.18)	(0.25)	(.28)
Other less attractive	3.21***	2.23***	2.30***
	(0.21)	(0.20)	(.22)
Other less attractive × F2F		1.39***	1.33***
		(0.27)	(.29)
Other less attractive × F2F High Stakes		1.55***	1.39***
		(0.42)	(.42)
Other less attractive × Anonymous		0.95*	.86*
		(0.50)	(.47)
Female			0.05
			(0.17)
Constant	−0.72***	−0.33	.51
	(0.19)	(0.20)	.035
Demographic controls	*N*	*N*	*Y*
Observations	399	399	376
Clusters	20	20	19
R-squared	0.436	0.450	0.472

Note: The table presents an OLS regression with the deviation of the assessment of the others' attractiveness rank from the actual rank as dependent variable. A deviation of zero indicates that participants are correct in their assessment. Negative values indicate that individuals assess their counterpart as less attractive than that person's actual attractiveness rank. Likewise, positive values indicate that individuals assess their counterpart as more attractive than that person's actual attractiveness rank. We include treatment dummies, a dummy for whether the other's actual rank is categorized as less attractive (ranks 7–10) and the interaction of that with the treatment variables as independent variables. The reference category is the judgment treatment. Female is a dummy variable coded as 1 if the participant was a woman and zero otherwise. Demographic controls include participants' age and ethnicity dummies. Clustered standard errors (at the session level) are reported in parenthesis. Robust standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

feedback to the less attractive individuals ($\chi^2(8) = 17.77$, $p = 0.023$). For the attractive individuals the two distributions do not differ ($\chi^2(8) = 8.62$, $p = 0.375$).

The results on the deviation between participant's actual ranks and the feedback provided to them in this treatment are almost identical to those observed in the F2F treatment with low stakes. For participants matched with attractive participants the average deviation is −0.23 (vs. −0.15 in F2F) while for those matched with the less attractive participants it is 3.55 (vs. 3.48 in F2F). As in the F2F treatment, when participants are matched with a less attractive participant, the average deviation is significantly larger in the F2F High Stakes treatment than in the Judgment treatment (MW test, $z = -4.15$, $p < 0.001$), while we do not find differences for participants matched with an attractive participant (MW test, $z = -0.26$, $p = 0.796$). These results are confirmed by the OLS regression reported in Table 2. The F2F High Stakes treatment dummy is significant and positive in column (1). Further, the interaction between the treatment dummy and the attractiveness dummy reported in columns (2) and (3) shows that on average the deviation is larger if the receiver is less attractive.

The reluctance to give honest feedback to less attractive individuals could be driven by the desire to protect one's own image. Alternatively, participants' unwillingness to hurt another person's feelings by giving negative feedback may be due to altruistic preferences. To distinguish between these two mechanisms, we study the effect of anonymous feedback. If people are reluctant to hurt other people's feelings, we expect similar behavior with both types of feedback. Instead, if participants (also) care about being identified as the person who is providing the negative feedback, we expect more honest feedback when their identity is kept anonymous.

We find that feedback in the Anonymous treatment is less positive than in the F2F treatments (see Fig. 2D). The average rank is similar to the average observed in the Judgment treatment (mean rank 4.84, $N = 99$). The distributions of ranks provided in the Anonymous and the Judgment treatment are not significantly different from each other ($\chi^2(9) = 7.07$, $p = 0.630$). The difference between the distribution of ranks in the Anonymous and the F2F treatment, however, is marginally significant ($\chi^2(8) = 14.98$, $p = 0.091$). Comparing the distributions of participants matched with an attractive and those matched with a less attractive counterpart separately yields insignificant results in both cases ($\chi^2(8) = 12.38$ and 7.38, $p \geq 0.135$). The difference between the distributions of ranks in the Anonymous and the F2F High Stakes treatment is not significant ($\chi^2(8) = 12.88$, $p = 0.168$).

The fraction of individuals who evaluated their partner as less-attractive is around three times higher in the Anonymous treatment than in the F2F or F2F High Stakes treatment (7% vs. 20%, $\chi^2(1) = 7.39$, $p = 0.007$). The average deviation between the feedback provided in stage three and the less attractive participants' actual rank is 3.48 in the F2F treatment, while it is only 2.58 in the Anonymous treatment. The difference between treatments is statistically significant (MW test, $z = 2.04$, $p = 0.042$). For the attractive individuals we find no significant difference; the average deviation is −0.15 in the

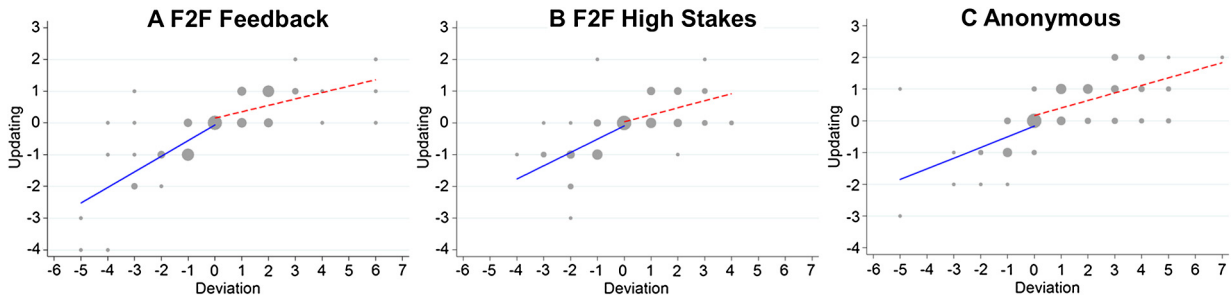

Fig. 3. Updating behavior after feedback. *Note:* The *x*-axis shows the deviation between the rank that participants receive as feedback and their initial self-evaluation (from Stage 2). Negative values indicate feedback that is more positive than individuals' self-evaluation while positive values indicate feedback that is negative compared to participants' initial self-evaluation. The *y*-axis shows the difference between the updated self-assessment in stage four and the self-assessment in stage one. Negative values indicate an update to a better rank while positive values indicate an update to a worse rank. The blue line displays the estimated OLS regression line for when participants receive positive feedback, while the red line displays the estimated OLS regression line for the cases in which participants receive negative feedback. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

F2F treatment and -0.61 in the Anonymous treatment (MW test, $z = 0.95$, $p = 0.340$). When we compare the deviations between the Anonymous and the Judgment treatment, we find that they do not differ for the attractive individuals (average deviation $= -0.33$ versus -0.61, MW test, $z = 0.61$, $p = 0.540$) and are marginally different for the less attractive participants (average deviation $= 1.9$ versus 2.58, MW test, $z = -1.73$, $p = 0.084$). Our results are confirmed by the OLS regression reported in Table 2. The Anonymous treatment dummy is insignificant in column (1). The interaction between the treatment dummy and individuals' actual rank in columns (2) and (3) is marginally significant, but the coefficient is smaller than that of the interaction between actual rank and the other feedback treatments.

Result 2: Face-to-face feedback is positively biased when assessing less attractive participants, irrespective of the size of the incentives. Anonymous assessment is less biased.

In conclusion, our results suggest that the reluctance to provide negative face-to-face feedback to less attractive individuals is mainly driven by individuals' unwillingness to be identified as the messenger of bad news. Participants provide more accurate negative feedback when the anonymity of the feedback provider is guaranteed.

3.3. Updating

We now investigate how participants react to the feedback they receive and whether they use it to update their own rank. Fig. 3 displays for all three feedback treatments the number of ranks by which participants update in stage four as a function of the number of ranks by which the feedback received deviates from participant's initial self-assessment. The OLS regression lines plotted in the figure display updating behavior as a function of this deviation, separately for cases in which the feedback received is positive and cases in which it is negative. As can be seen in Figs. 3A and 3B for the two F2F treatments, the regression line tends to be less steep when participants receive negative face-to-face feedback, suggesting that these participants update to a lower extent than participants who receive positive feedback. In Fig. 3C for the Anonymous treatment, the slopes are however similar.

The OLS regression reported in Table 3 reveals that in all three treatments the larger the difference between participants' self-evaluation and the feedback they receive, the more they update. For F2F and F2F High Stakes this is the case both if participants receive positive feedback (F2F: $\beta = .54$, $p = 0.025$, F2F High Stakes: $\beta = .44$, $p < 0.006$) and if they receive negative feedback (F2F: $\beta = .23$, $p = 0.021$, F2F High Stakes: $\beta = .25$, $p =< 0.091$). In the Anonymous treatment, we only find a significant effect when participants receive negative feedback (positive feedback: $\beta = .41$, $p = 0.136$, negative feedback: $\beta = .28$, $p = 0.002$). Although the coefficients are more than double the size for positive than for negative feedback, the differences between the coefficients are not statistically significant in any of the treatments ($p >= 0.62$), i.e. we cannot reject the hypothesis that participants update at the same extent when feedback is positive and when it is negative. This finding is different from the finding of Eil and Rao (2011) and Möbius et al. (2014), who show a significantly greater responsiveness to "good news" rather than "bad news". However, if we analyze the data without clustering the standard error at the session level, the difference becomes significant. For additional analyses on updating behavior see Appendix A.

Result 3: A majority of participants update their self-assessment in the direction of the feedback they receive both when feedback is positive and when it is negative.

Feedback should help individuals make more precise guesses. In order to explore whether receiving feedback helps the recipient, we investigate how individuals' precision in their self-assessment changes after feedback (stage four) as compared to before feedback (stage two).

Table 3
Degree of updating by treatment.

Extent of updating	(1) F2F	(2) F2F	(3) F2F High Stakes	(4) F2F High Stakes	(5) Anonymous	(6) Anonymous
Self-message received (pos.)	0.54**	0.53**	0.44***	0.44***	0.41	0.41
	(0.15)	(0.15)	(0.08)	(0.08)	(0.22)	(0.22)
Self-message received (neg.)	0.23**	0.24**	0.25*	0.25*	0.28***	0.28***
	(0.06)	(0.06)	(0.11)	(0.11)	(0.04)	(0.04)
Attractiveness dummy		0.13		0.09		−0.04
		(0.24)		(0.10)		(0.07)
Constant	0.08	−0.01	−0.04	−0.09	0.03	0.05
	(0.15)	(0.16)	(0.14)	(0.16)	(0.13)	(0.15)
Observations	98	98	99	99	99	99
Clusters	20	20	20	20	20	20
R-squared	0.622	0.626	0.455	0.457	0.493	0.493

Notes: The table presents OLS estimates. Extent of updating indicates by how much individuals update in stage four as compared to their original self-evaluation provided in stage two: negative values indicate that individuals updated to a better rank while positive values indicate that individuals updated to a worse rank. Self-message received (pos.) is a variable that indicates the difference between participants' self-evaluation in stage two and the message they receive from others when feedback is positive (as compared to participants' self-evaluation), and is coded as 0 otherwise. Self-message received (neg.) is a variable that indicates the difference between participants' self-evaluation and the message they receive from others when feedback is negative, and is coded as 0 otherwise. Positive (negative) feedback means that the feedback indicates a better (worse) rank than the self-evaluation from Stage two. Attractiveness is a dummy coded as 1 if individuals are attractive (ranks 1–6). Clustered standard errors (at the session level) are reported in parenthesis. Robust standard errors in parentheses *** $p < 0.1$, ** $p < 0.05$, * $p < 0.01$.

In the two F2F treatments we observe that feedback does not improve the accuracy of the assessment neither for the attractive nor the less attractive participants. In the F2F treatment for the attractive participants the deviation between individuals' self-assessment and their actual rank changes from 0.07 to 0.08 (WS test, $z = 0.26$, $p = 0.795$) while for the less attractive individuals it changes from 3.66 to 3.98 (WS test, $z = -0.73$, $p = 0.467$). In the F2F High Stakes treatment the deviation for attractive participants does also not significantly change after feedback (0.17 to 0.22, WS test, $z = 0.37$, $p = 0.709$), while for the less attractive participants it even becomes significantly larger, changing from 3.38 to 3.65 (WS test, $z = -2.18$, $p = 0.029$).

Given that the feedback in the Anonymous treatment is more honest, we find that it is beneficial to the participants, as it helps them correct their biased self-evaluation. The average difference between participants' self-evaluation and their actual rank significantly decreases from 3.7 to 3.4 (WS test, $z = 2.19$, $p = 0.028$) for the less attractive participants, while for the attractive ones it decreases from 0.43 to 0.2 (WS test, $z = 1.98$, $p = 0.048$).

Result 4: Only anonymous feedback improves the accuracy of self-assessment.

4. Concluding remarks

Are there "must-lie" situations? We show that when asked to give negative face-to-face feedback people prefer to lie rather than provide an honest assessment. This aversion to providing negative feedback is costly because it slows down the learning process and possibly encourages superfluous investments of resources such as time or money. Our results suggest that especially in cases where an individual's own perception is the furthest away from the perception of others, a lack of honest feedback leads to greater distortion of self-perception and a worse outcome than receiving no feedback. Increasing the stakes fivefold does not help to outweigh the costs associated with the open delivery of negative feedback.

We focus on attractiveness as a proxy for traits or characteristics for which receiving negative signals is costly. In many cases, the scope for adjustments after receiving negative feedback on attractiveness is limited, and thus the opportunity costs associated with receiving overly positive feedback are low. In our experiment, however, an improvement in participants' self-assessment could increase their earnings. While attractiveness is a fairly sensitive attribute, feedback on other characteristics may generally be less hurtful. Future work should explore the generalizability of our findings to other contexts in which subjective and sensitive attributes are relevant, and where negative but honest feedback can be instrumental for achieving better outcomes.

A bias in the provision of feedback in organizations, for example, could have strong negative effects on the efficiency of team work, as this critically hinges on honest communication. In situations in which giving negative feedback to a co-member of the organization may reflect badly on the sender, the organization can suffer from a particularly slow learning process. The observation that the aversion to giving negative feedback is reduced when it is done anonymously suggests that guaranteeing the anonymity of the feedback provider is important for the accuracy and informational value of feedback. In an organizational context, our results suggest that in addition to having an open feedback round, organizations could consider installing a "feedback box" in which staff can anonymously provide feedback to their colleagues or bosses.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.geb.2017.01.008.

References

Akerlof, R., Kranton, R., 2000. Identity and economics. Quart. J. Econ. 115, 715–753.

Ariely, D., Bracha, A., Meier, S., 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. Amer. Econ. Rev. 99, 544–555.

Balafoutas, L., Kerschbamer, R., Sutter, M., 2012. Distributional preferences and competitive behaviour. J. Econ. Behav. Organ. 83 (1), 125–135.

Burks, S.V., Carpenter, J.P., Goette, L., Rustichini, A., 2013. Overconfidence and social signaling. Rev. Econ. Stud. 80 (3), 949–983.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.

Crawford, V., Sobel, J., 1982. Strategic information transmission. Econometrica 50 (6), 1431–1451.

Dreber, A., Johannesson, M., 2008. Gender differences in deception. Econ. Letters 99 (1), 197–199.

Eil, D., Rao, J.M., 2011. The good news-bad news effect: asymmetric processing of objective information about yourself. Amer. Econ. J. Microecon. 3 (2), 114–138.

Englmaier, F., 2006. A brief survey on overconfidence. In: Satish, D. (Ed.), Behavioral Finance – an Introduction. ICFAI University Press.

Erat, S., Gneezy, U., 2012. White lies. Manage. Sci. 58 (4), 723–733.

Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise – an experimental study on cheating. J. Eur. Econ. Assoc. 11, 525–547.

Freud, S., 1991. On Metapsychology. Penguin Freud Library, vol. II, pp. 454–455.

Gneezy, U., 2005. Deception: the role of consequences. Amer. Econ. Rev. 95 (1), 384–394.

Gneezy, U., Saccardo, S., Serra-Garcia, M., van Veldhuizen, R., 2016. Bribing the Self. Mimeo.

Levitt, S., Snyder, C., 1997. Is no news bad news? Information transmission and the role of “early warning” in the principal-agent model. RAND J. Econ. 28 (4), 641–661.

Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: a theory of self-concept maintenance. J. Marketing Res. 45 (6), 633–644.

Möbius, M., Rosenblat, T., 2006. Why beauty matters. Amer. Econ. Rev. 96, 222–235.

Möbius, M., Niederle, M., Niehaus, P., Rosenblat, T., 2014. Managing Self-Confidence. Working paper.

Moore, D., Healy, P., 2008. The trouble with overconfidence. Psychol. Rev. 115 (2), 502–517.

Morris, S., 2001. Political correctness. J. Polit. Economy 109 (2), 231–265.

Olszewski, W., 2004. Informal communication. J. Econ. Theory 117 (2), 180–200.

Ottaviani, M., Sørensen, P., 2005a. Professional advice. J. Econ. Theory 126 (1), 120–142.

Ottaviani, M., Sørensen, P., 2005b. Reputational cheap talk. RAND J. Econ. 37 (1), 155–175.

Prendergast, C., 1993. A theory of “yes men”. Amer. Econ. Rev. 83 (4), 757–770.

Solnick, S., Schweitzer, M., 1999. The influence of physical attractiveness and gender on ultimatum game decisions. Org. Behav. Hum. Decis. Process. 79 (3), 199–215.

Sutter, M., 2009. Deception through telling the truth? Experimental evidence from individuals and teams. Econ. J. 119 (534), 47–60.

Svenson, Ola, 1981. Are we all less risky and more skillful than our fellow drivers? Acta Psychol. 47, 143–148.