A newly detected bias in self-evaluation

Guillaume Deffuant, Thibaut Roubin, Armelle Nugier, Serge Guimond

A newly detected bias in self-evaluation

Guillaume Deffuant[1,2], Thibaut Roubin[1], Armelle Nugier[2], Serge Guimond[2]

[1]Université Clermont Auvergne, INRAE LISC,

[2] Université Clermont Auvergne, LAPSCO

Abstract

The widely observed positive bias in self-evaluation is mainly explained by the self-enhancement tendency which minimizes negative feedbacks and emphasizes positive ones. Recent agent based simulations suggest that a positive bias also emerges if the sensitivity to feedbacks decreases when the self-evaluation increases. We describe a pilot experiment (N = 220) and a larger experiment (N = 1509) aiming at detecting such a bias. The results confirm that, especially when the self-evaluation is high, the sensitivity to feedbacks tends to decrease when self-evaluation increases and this generates a specific positive bias.

A newly detected bias in self-evaluation

Introduction

People overrate themselves. On average, people say that they are "above average" in skill, over-estimate the likelihood that they will engage in desirable behaviors and achieve favorable outcomes, furnish overly optimistic estimates of when they will complete future projects, and reach judgments with too much confidence. This quotation from Dunning, Heath, and Suls (2004) is part of their review of numerous evidences of a positive bias in self-evaluation. They report in particular evidence of overoptimism or overconfidence in judgement and predictions, for instance about the duration of romantic relationship (Epley & Dunning, 2000) or the ability to complete a task (Dunning & Story, 1991) or about forecasting events in general (Buchler, Griffin, & Ross, 1991; Fischhoff, Slovic, & Lichtenstein, 1977; Griffin, Dunning, & Ross, 1990; Vallone, Griffin, Lin, & Ross, 1990).

Many psychological mechanisms are considered as potential explanations of these observations. In particular, informational deficits such as neglecting situational details, possible scenarios or background circumstances could be responsible to prediction overconfidence (Griffin et al., 1990) and failure to appreciate the effect of negative emotions (fear, anxiety and embarrassment) are proposed explanations.

A related and more general explanation is self-enhancement: most of us tend to seek out and accept positive feedbacks about themselves and avoid or reject negative ones. When we receive a negative feedback, we often tend to decrease our evaluation of the source of the feedback and thus we decrease its impact (Campbell & Sedikides, 1999). As a result, on average, negative feedbacks tend to have a lower impact than positive ones, which leads to self-overestimation (Moreland & Sweeney, 1984). According to this explanation, the self overestimation is a consequence of self-enhancement.

This paper considers a different type of positive bias in self-evaluation that can emerge without self-enhancement. This bias has been first observed in an agent based model (Deffuant, Bertazzi, & Huet, 2018). Indeed, the agents of this model increase

their self-evaluation without favouring positive feedbacks over negative ones, when they are submitted to a sequence of positive and negative experiences of same intensity. Deffuant et al. (2018) show that this bias is due to the decreasing sensitivity of the self-evaluation to the feedbacks (whether positive or negative) when the self-evaluation increases. In other words, in the considered agent model, when the self-evaluation is high, feedbacks (positive or negative) have a lower impact than when the self-evaluation is low. As shown in this paper, the positive bias observed after feedback fluctuations is a mathematical consequence of the decreasing sensitivity to the feedback.

The present paper reports experimental evidence that, like in the agent model, humans show a decreasing sensitivity to feedbacks, when their self-evaluation increases, and this induces a positive bias in self-evaluation. The purpose of the paper is also to experimentally check the hypotheses of the agent based model, an approach which is strongly recommended by Flache et al. (2017). Actually, in some respect, this paper goes beyond this recommendation as it also checks the consequence of these hypotheses that model simulations revealed.

The next section presents our hypothesis in more details, the experimental setting and the method used for treating the results. The following section reports the results of the experiments. The final section of the paper wraps up the results and proposes a discussion about them.

Material and method

Model and hypotheses

This section presents the model of self-evaluation modification when receiving feedbacks, derived from the agent model of Deffuant et al. (2018). It shows how a positive bias emerges from series of alternate positive and negative feedbacks, if the sensitivity to feedback decreases when the self-evaluation increases. Then it extends the approach to a model including self-enhancement.

Positive bias from decreasing sensitivity without self-enhancement. Consider an agent with self-evaluation a_t at time t when receiving feedback f_t (i.e. an

evaluation coming from an outside source). The main hypotheses from Deffuant et al. (2018) are:

- the change of self-evaluation due to this feedback is proportional to the difference between the feedback and the self-evaluation;
- Moreover, the coefficient of proportionality decreases with a_t.

These hypotheses are thus expressed by equation 1:

$$a_{t+1} - a_t = h(a_t)(f_t - a_t), \tag{1}$$

where $h(a_t)$ is a positive and decreasing function (after averaging possible random fluctuations). Note that, if function h is derivable, its derivative h' is such that: $h'(a_t) < 0$ for all a_t.

According to this model, for the same difference between feedbacks and self-evaluations (whether positive or negative), agents with a high self-evaluation can less easily be influenced than agents with a low self-evaluation. An argument for this model is that agents with a high self-evaluation tend to be more confident and thus less prone to change their mind.

Let $f_t - a_t$ be the intensity of feedback f_t. We say that a feedback is positive when its intensity is positive and negative otherwise. We show now that the previous model generates a positive bias when receiving a series of feedbacks of opposite intensities. We consider the simple example of an agent receiving two consecutive feedbacks of opposite intensities $\pm\delta$.

Assume that the agent starts with self-evaluation a_1 and receives first the positive feed-back $f_1 = a_1 + \delta$. Applying equation 1, the self-evaluation of the agent becomes a_2 with:

$$a_2 = a_1 + h(a_1)\delta. \tag{2}$$

Then the agent receives the negative feedback $f_2 = a_2 - \delta$ and its self-evaluation a_3 becomes:

$$a_3 = a_2 - h(a_2)\delta. \tag{3}$$

The difference of self-evaluation between before and after receiving the couple of feedbacks is:

$$a_3 - a_1 = a_1 + h(a_1)\delta - h(a_2)\delta - a_1 = (h(a_1) - h(a_2))\delta. \tag{4}$$

As we assume that at any time t, $h(a_t) > 0$, we have $a_1 < a_2$ and, if h is decreasing, we have: $h(a_1) - h(a_2) > 0$, hence $a_3 - a_1 > 0$.

Now, if we invert the order of the feedbacks ($f_1 = a_1 - \delta$ and $f_2 = a_2 + \delta$), we have:

$$a_3 - a_1 = (h(a_2) - h(a_1))\delta. \tag{5}$$

Now $a_2 < a_1$, therefore again, because h is decreasing $a_3 - a_1 > 0$.

The mathematical analysis shows that after receiving two feedbacks of opposite intensities, the self-evaluation tends to increase. Note that this increase takes place without self-enhancement, because in equation 1, the amplitude of the reaction to the positive feedback $a_t + \delta$ and to the negative feedback $a_t - \delta$ is the same: $|h(a_t)\delta|$.

However, the mathematical analysis also suggests that this positive bias is rather small. Indeed, assuming that h is derivable and δ is relatively small, developing $h(a_2)$ at the first order, we get:

$$h(a_2) \approx h(a_1) + h'(a_1)h(a_1)\delta, \text{ if } f_1 = a_1 + \delta; \tag{6}$$

$$h(a_2) \approx h(a_1) - h'(a_1)h(a_1)\delta, \text{ if } f_1 = a_1 - \delta. \tag{7}$$

Therefore, for both sequences of feedbacks we get:

$$S(a_1) = a_3 - a_1 \approx -h'(a_1)h(a_1)\delta^2. \tag{8}$$

This positive bias is thus expected to be of the second order of the intensity of the feedback, hence rather small. This could explain why, if this bias exists in humans, it has not been noticed yet.

Moreover, Deffuant et al. (2018) show that the bias appears also when δ is uniformly drawn in interval $[-\delta_M, \delta_M]$, hence it is not restricted to the very specific case of feedbacks of opposite intensities. We focus on this case in this paper because it makes an experiment simpler to define.

Positive bias from decreasing sensitivity with self-enhancement. In the framework of this model, self-enhancement takes place when assuming that the sensitivity $h_p(a_t)$ to positive and $h_n(a_t)$ to negative feedbacks are different :

$$a_{t+1} - a_t = h_p(a_t)\delta, \text{ if } f_t = a_t + \delta, \tag{9}$$

$$a_{t+1} - a_t = -h_n(a_t)\delta, \text{ if } f_t = a_t - \delta. \tag{10}$$

Considering feedbacks of intensity $\pm\delta$, the bias of self-enhancement $E(a)$ at a given self-evaluation a can be expressed as the difference between the reaction to the positive feedback $f_p = a + \delta$ and the reaction to the negative feedback $f_n = a - \delta$:

$$E(a) = (h_p(a) - h_n(a))\delta. \tag{11}$$

Now, assume that the agent's self-evaluation is a_1 and that the agent receives a positive and then a negative feedback. Repeating the previous calculations, we get:

$$a_2 = a_1 + h_p(a_1)\delta, \tag{12}$$

$$a_3 = a_2 - h_n(a_2)\delta. \tag{13}$$

The total bias $T(a_1)$ from these successive feedbacks is:

$$T(a_1) = a_3 - a_1 \tag{14}$$

$$= (h_p(a_1) - h_n(a_1))\delta - h'_n(a_1)h_p(a_1)\delta^2. \tag{15}$$

We recognise the self-enhancement bias in the first term and the bias from decreasing sensitivity in the second term:

$$S(a_1) = -h'_n(a_1)h_p(a_1)\delta^2. \tag{16}$$

This value is positive when $h'_n(a_1)$ is negative and we have:

$$T(a_1) = E(a_1) + S(a_1). \tag{17}$$

Moreover, if we have a series of 2 positive and 2 negative feedbacks in a random order (as it will be the case in the experiment), the average bias from decreasing sensitivity is:

$$S(a) = \frac{1}{4}\left(-h'_n(a)h_p(a) - h'_p(a)h_n(a) - h'_p(a)h_p(a) - h'_n(a)h_n(a)\right)\delta^2, \tag{18}$$

$$S(a) = -h'_m(a)h_m(a)\delta^2, \tag{19}$$

where h_m is the average of h_p and h_n: $h_m(a) = \frac{1}{2}(h_p(a) + h_n(a))$. In the following experiments, we derive average values of functions h_n and h_p from data collected on several participants and then we evaluate the biases from self-enhancement and decreasing sensitivity using the above formulas.

Experiments

The experiment design has been approved by the committee of ethics from Clermont Auvergne Université. The participants were recruited by a specialised company. They live in various regions of France and answer to an online questionnaire. In the core of the experiments, the participants receive a series of 4 feedbacks, two positive, two negative, of same intensity in absolute value, starting from different self-evaluations. The main objective is to collect data about the sensitivities to positive and negative feedbacks (functions h_p and h_n in the model). We first performed a pilot experiment involving 220 participants and then the main experiment involving 1509 participants. The results of the pilot experiment led us to make minor changes in the protocol of the main experiment.

Online questionnaire. The questionnaire of both experiments includes the following steps:

- The participants are requested to assess the size of a colored surface in 3 different 2D images. An example of image is shown on figure 1.

- The participants are told that the experimenters can compute exactly their error of surface assessment on these three images and can do the same for a large number of other people who already performed the task. Moreover, the participants are told that the experimenters gathered at random 6 different groups (G_0 to G_5) of 100 people whose performance at the task is known and that the performance of the participant will be compared to the one of the members of these groups. This comparison provides an evaluation, between 1 and 100, of the participant with respect to the group. We tested two evaluation scales: rank and score which are described further.

- The participants are given the evaluation f_0 of their error within G_0, the first group of 100 persons. We call f_0 the anchor because it is the initial reference evaluation for the participant. This anchor is actually defined by the experimenters in a way that is described further.

- Given their anchor, the participants are asked to express their expected evaluation in the second group of 100 people (G_1). We interpret this evaluation a_1 as the first self-evaluation.

- The feedback f_1 is presented as the evaluation of the participant in group G_1, computed by comparing errors at the task. It is actually defined automatically as:

$$f_1 = a_1 \pm \delta \pm \epsilon, \tag{20}$$

  where $\delta = 13$ and $\epsilon = 1$. The addition of the small variation $\pm\epsilon$ aims at avoiding to produce too regular series of feedbacks that could undermine the confidence of the participant in the reality of the feedback as an actual evaluation within the group.

- The participants are asked their expected evaluation a_2 in group G_2. They are requested to express this evaluation between their previous expectation a_1 and the feedback f_1 that they just received.

- The same process is repeated again three times, with feedbacks f_2, f_3 and f_4 that are presented as the evaluation of the participant in groups G_2, G_3 and G_4, and requesting the participant's expected evaluations a_3, a_4 and a_5 in groups G_3, G_4 and G_5. Actually, each time, the feedbacks are computed as:

$$f_t = a_t \pm \delta \pm \epsilon, \tag{21}$$

  where a_t is the expected evaluation of the participant in group G_t given the last feedback f_{t-1} which is (allegedly) their evaluation in group G_{t-1}.

- Finally, the participants are asked if they believed that the feedbacks were really the evaluation of their error in real groups of 100 persons or if they believed that these feedbacks were manipulated by the experimenters. The participants are requested to rate their belief between 1 (the feedbacks are fake) to 10 (the feedbacks are real). In the following, we call this answer: "trust in feedback" or sometimes simply "trust" of the participant.

The sequence of positive and negative feedbacks is chosen at random in the six possible sequences that contain two positive and two negative feed-backs (see table in Fig: 1). However, in some cases, when the expected evaluation a_t is close to the limit 1 or 100, the chosen feedback would leave the [1,100] interval. In these cases, the feedback is truncated in order to remain in [1,100]. This might lead to some sequences where the positive and negative feedbacks are not balanced.

Finally, the experiment also includes a questionnaire evaluating the self-esteem of the participants using Rosenberg's scale (Vallières & Vallerand, 1990).

Protocols. The main variables of the experimental protocol are:

- The anchor (first feedback that influences the first self-evaluation of the participant), which is an integer in [1,100];

- The type of evaluation which can be:

  - Rank: The rank of the participant's performance at the task within the group of 100 persons. 1 is the best rank, 101 is the worst:

f_0	f_1	f_2	f_3
+	+	-	-
+	-	+	-
+	-	-	+
-	+	+	-
-	+	-	+
-	-	+	+

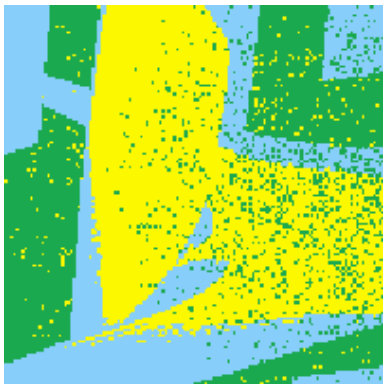

Figure 1. Left panel: the 6 possible sequences of 4 feedbacks. Right panel: Example of image used in the task. The participants are requested to evaluate the percentage of surface in green on three images similar to this one.

  - Score: The number of persons in the group whose performance at the task is worse than the one of the participant. 100 is the best score, 0 is the worst.

The protocols of the pilot and main experiments are different:

- In the pilot:
  - The anchor is randomly drawn between 5 and 95;
  - The evaluation is by rank for all participants.
- In the main experiment:
  - One third of the participants were given a low anchor (randomly chosen in [5, 40]) two third were given a high anchor (randomly chosen in [60, 95]). Indeed, the pilot experiment suggested that sensitivity to feedbacks decreases more strongly when the anchor is high and we wanted to check this on more data.
  - For half the participants, the evaluation is by rank like in the pilot experiment and for the other half, the evaluation is by score. Indeed, the pilot experiment results suggest that, when the anchor is low, the sensitivity to the feedback increases. We wanted to check if it was coming from the evaluation by rank.

Result treatment

The data of both the pilot and the main experiment is available at `https://osf.io/c3kt6/`. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. First, in order to simplify the presentation of the results, we only use the evaluation by score. Hence the first treatment is to transform any rank r whether allegedly computed by performance comparison in a group (for feedbacks) or expected by the participant (for self-evaluations) into $100 - r$.

The experiments yield a set of triples including self-evaluation at t, feedback at t, self-evaluation at $t+1$, denoted by $(a_t^i, f_t^i, a_{t+1}^i)$, the exponent i designating the participant to the experiment and $t \in \{1, 2, 3, 4\}$ being the index of the successive feedbacks and self-evaluations for this participant. The treatments are made on a subset A of all the triples collected from an experiment.

As in the model we assume that the agents have a complete trust in the feedbacks, in the result treatment, we particularly focus on participants who have a high trust in the feedback.

Regressions of self-evaluation change by self-evaluation. Our first aim is to check the hypothesis that sensitivities $h_p(a_t)$ to positive feedbacks and $h_n(a_t)$ to negative feedbacks are decreasing. According to equations 12 and 13, we have:

$$\frac{a_{t+1} - a_t}{f_t - a_t} = h_p(a_t) \text{ for positive feedbacks;} \tag{22}$$

$$\frac{a_{t+1} - a_t}{f_t - a_t} = h_n(a_t) \text{ for negative feedbacks.} \tag{23}$$

Notice that $f_t - a_t = \pm\delta \pm \epsilon$ and the questionnaire limits the participant's choice of a_t so that in all cases $a_{t+1} \in [a_t, f_t]$ if f_t is positive and $a_{t+1} \in [f_t, a_t]$ if f_t is negative and $0 \leq a_t \leq 100$. We used the variable $a_t/100$ instead of a_t in the regression as it is more usual to perform regressions on values in [0,1].

Let the regression of $\frac{a_{t+1}-a_t}{f_t-a_t}$ as a function of $\frac{a_t}{100}$ for the positive feedbacks be the linear model $c_p\frac{a_t}{100} + b_p$. It is a linear approximation of $h_p(a_t)$:

$$c_p\frac{a_t}{100} + b_p \approx h_p(a_t). \tag{24}$$

Similarly, let the linear model $c_n \frac{a_t}{100} + b_n$ be the same regression for the negative feedbacks. This linear model is an approximation of the sensitivity to negative feedbacks:

$$c_n \frac{a_t}{100} + b_n \approx h_n(a_t). \tag{25}$$

The sign of coefficients c_p and c_n indicates if these approximations are increasing or decreasing.

Total bias T. This measure is the average change of self-evaluation after the series of 4 feedbacks, divided by twice the average difference $|f_t - a_t|$ in the series, for all the participants i in the considered sample of results A (p_A being the number of triples in this sample).

$$T(A) = \frac{1}{p_A} \sum_i \frac{a_5^i - a_1^i}{\frac{1}{2} \sum_{t \in \{1,\dots,4\}} |f_t^i - a_t^i|}. \tag{26}$$

Indeed, $T(A)$ is the average difference between the last self-evaluation of the series of two positive and two negative feedbacks (a_5) and the first one (a_1). We divide this value by the half the sum of $|f_t^i - a_t^i|$ for $t = 1, .., 4$ because we evaluate the average change for a series of two opposite feedbacks as a percentage of difference $|f_t^i - a_t^i| = \delta \pm \epsilon$. Therefore, $T(A)$ is the average change of self-evaluation after a series of two opposite feedbacks, as a percentage of $\delta \pm \epsilon$. Moreover, $T(A)$ is the sum of the self-enhancement bias and the bias from decreasing sensitivity (see equation 17).

Self-enhancement bias E. The average self-enhancement bias is the average on A of the self-enhancement for a_t^i as defined by see equation 11, i.e. the difference between the reaction to a positive and to a negative feedback at a_t^i.

$$E(A) = \frac{1}{4p_A} \sum_{i,t} (c_p - c_n) \frac{a_t^i}{100} + b_p - b_n, \tag{27}$$

where p_A is the number of participants in sample A.

The self-enhancement bias E can be seen as the average change of self-evaluation after a sequence of two opposite feedbacks, without taking into account the variation of

sensitivity to the feedback. We express this change as a percentage (here since $|f_t - a_t| = \delta \pm \epsilon$, it is roughly a percentage of δ).

Note that the self-enhancement bias can be negative. In this case, the participants are, on average, more sensitive to the negative than to the positive feedbacks.

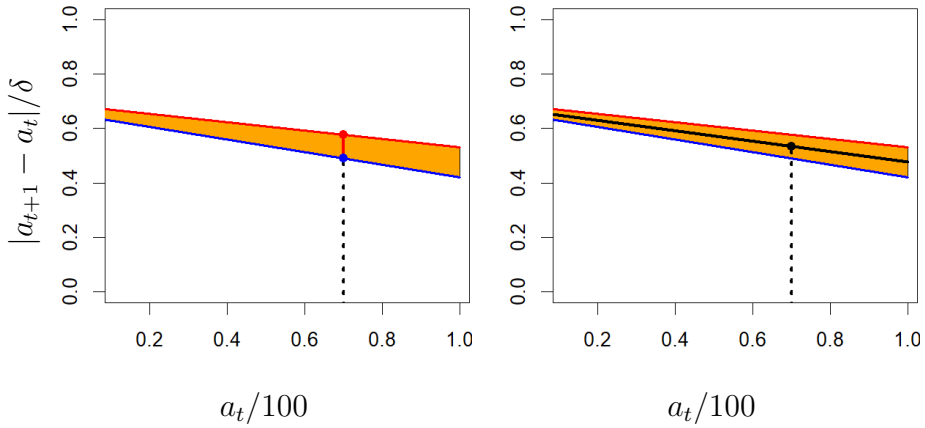

Figure 2. Illustration of the computation of biases from self-enhancement and sensitivity. The self-evaluation $a_t/100 = 0.7$ is shown by the dotted line. In both panels, the red and blue lines are the linear approximation of the sensitivity to respectively the positive and negative feedbacks. On the left panel, the self-enhancement for a_t is the distance between the projections of point $(a_t/100, 0)$ on the blue and red lines respectively represented by a blue and a red point. On the right panel, the bias from sensitivity is $-c_m(c_m(a_t/100) + b_m)\delta$, where $c_m x + b_m = 0$ is the equation of the medium line between the red and blue lines.

Theoretical bias from sensitivity of feedbacks S_t. This measure is the theoretical average change of self-evaluation due to the decreasing sensitivity. Following formula 19, this measure is:

$$S_t(A) = \frac{1}{4p_A} \sum_{i,t} -c_m \left(c_m \frac{a_t^i}{100} + b_m \right) \left(f_t^i - a_t^i \right), \tag{28}$$

where $c_m = \frac{c_p + c_n}{2}$ and $b_m = \frac{b_p + b_n}{2}$. Note that we divided equation 19 by $f_t^i - a_t^i$ (which is $\pm\delta \pm \epsilon$), so that this measure is expressed as a percentage like self-enhancement and total biases. The difference between total and self-enhancement biases is the bias from

sensitivity and it should be close to the theoretical value:

$$S(A) = T(A) - E(A) \approx S_t(A). \tag{29}$$

Figure 2 illustrates the computation of the biases from self-enhancement and sensitivity.

Results

Pilot experiment

The experiment involves 220 participants (103 male 117 female, age between 17 and 75). After removing the triples with truncated feedbacks, there are 772 triples (a_t, f_t, a_{t+1}) left.

First, we expect that the behaviour of the participants depends on their trust in the reality of the feedback. Indeed, the behaviour of the participants with low trust is likely to be influenced by their imagination of a manipulation from the experimenters. Therefore, in the following, we always contrast the results from samples of participants of high (≥ 7) to low (≤ 3) trust. We are particularly interested in the data from participants with a high trust, as it corresponds to the assumption of the agent model.

- For all participants of low (≤ 3) trust, the bias from sensitivity is $S = -0.01\%$, its theoretical approximation is $S_t = 0.08\%$, the self-enhancement bias is 4.51% and the total bias is 4.52%.

- For all participants of high (> 7) trust,the bias from sensitivity is $S = 0.42\%$, its theoretical approximation is $S_t = 0.77\%$, the self-enhancement bias is $E = 11.43\%$ and the total bias for $T = 11.85\%$. The values of the regression coefficients are $c_p = -0.03$ and $c_n = -0.18$. Hence, we detect a slight bias from sensitivity in this second sample, but the regression coefficients are not statistically significant.

These first results led us to differentiate between participants of low and high anchors. Indeed, in the model hypothesis (Deffuant et al., 2018; Deffuant, Carletti, & Huet, 2013), function $h(a_t)$ is non-linear, which means that the slope is probably stronger in a particular region of a_t.

Figure 3 shows the different measures of biases and the regression coefficients computed for 8 different samples, distinguishing low (≤ 50) anchor and high (≥ 50) anchor and for each case, two samples of low (≤ 2 and ≤ 3) trust in feedbacks and two samples of high (≥ 7 and ≥ 8) trust in feedbacks. The condition on trust is specified by the x-axis of each panel and the vertical dotted line separates the samples of low (on the left) and high (on the right) trust in feedbacks. On average the considered samples contain 125.9 triples (a_t, f_t, a_{t+1}) with a standard deviation of 27.3.

- The panels of the top row show the measure of self-enhancement bias E (bars) and its difference with the total bias T (error bar), both expressed as a percentage of the feedback intensity δ. Remember that the difference $T - E$ is a direct measure of the bias from sensitivity S. The self-enhancement bias is of 10 % of δ or more for all samples except for high anchor and high trust, where it is is much smaller. The bias from sensitivity S is of the same order (2 to 4 % of δ) for all the samples. For the samples of high anchor and high trust the bias from sensitivity is larger or similar to the self-enhancement bias.

- The panels of the middle row show the bias from sensitivity S (bars) and its difference with the theoretical measure of the bias from sensitivity S_t (error bars) obtained using equation 28. The highest positive bias from sensitivity is measured for high trust and high anchor. Note that S is negative for high anchor and low trust, and for low anchor and high trust. The average difference $|S - S_t|$ on the 8 considered samples is 0.44.

- The bottom row of panels represents the value of the regression coefficient c_n for negative feedbacks (bars) and the regression coefficient c_p for positive feedbacks (error bars). For low anchor and low trust, coefficients c_n are negative while coefficients c_p are positive. The bias from sensitivity is of the sign of $-(c_n + c_p)$.

Figure 4 shows examples of samples and regressions for positive (red line) and negative (blue line) feedbacks for all combinations of low (≤ 50), high (≥ 50) anchors with low (≤ 3) and high (≥ 7) trust. The orange area between the lines represents

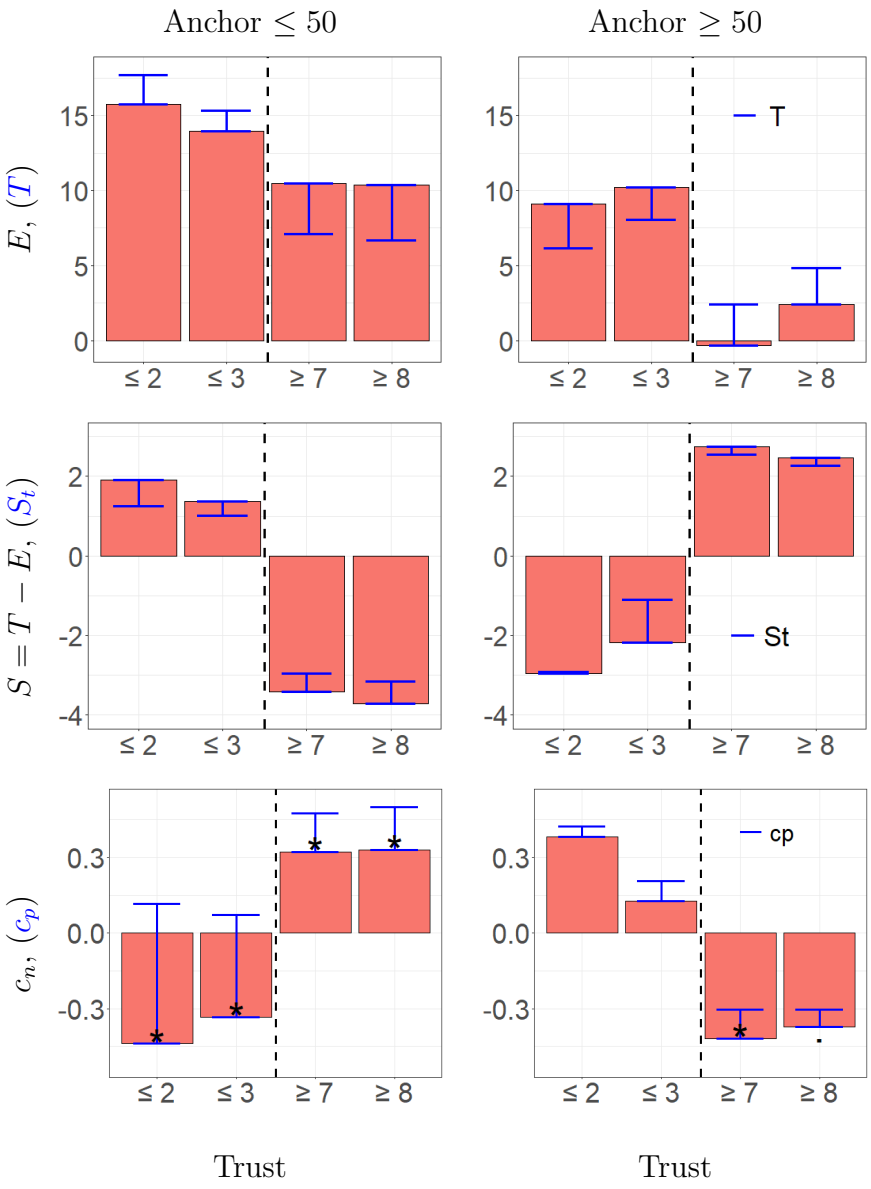

Figure 3. Measures of biases an regressions coefficients on pilot data. Top row panels: Self-enhancement bias E, with error bars showing the difference to the total bias (T). Middle row panels: Bias from sensitivity $S = T - E$, with error bars showing the difference to the theoretical bias from sensitivity S_t. Bottom row panels: Regression coefficients c_n for negative feedbacks, error bars showing the difference with c_p, the regression coefficient for positive feedbacks. The x-axis in each panel shows the condition of trust in feedbacks.

positive self-enhancement and the blue area represents negative self-enhancement. In all cases, the self-enhancement increases when the self-evaluation (a_t) increases. On the top

right panel, the self-enhancement is slightly negative.

Overall, the pilot experiment suggests the existence of a positive bias from sensitivity for high anchor and high trust, but also the existence of a negative bias from sensitivity for low anchor and high trust. We made the hypothesis that this negative bias from an increasing sensitivity could be related to the choice of the evaluation by rank. Indeed, when the self-evaluation increases, the rank r decreases and the constant intensity δ of the feedback increases relatively (i.e. δ/r increases). In order to check this hypothesis, we also tested the evaluation by score in the main experiment. Moreover, as we want to check the significance of the regression coefficient in the region of high anchors, we decided to collect more data in this region in the main experiment.

Main experiment

The experiment involves 1509 participants (706 male, 803 female, age between 17 and 79). After removing the triples with truncated feedbacks, there are 5472 triples (a_t, f_t, a_{t+1}) left. We directly present the results distinguishing rank and score evaluations. Indeed, these results show that self-enhancement biases depend heavily on the evaluation type and it appeared irrelevant to mix them. Then we show some results about the effect of self-esteem.

Results distinguishing low and high anchors. Figure 5 shows the different measures of biases and the regression coefficients for high and low anchors, like in figure 3, distinguishing between evaluations by score and by rank. The average sample size is 220.5 (s.d. 46.9) for low anchor and 668.5 (s.d. 158.2) for high anchor.

- The panels in the top row show that self-enhancement is positive for evaluations by rank and is negative or close to zero for evaluations by score. For evaluations by rank, self-enhancement is significantly positive for high anchor and high trust which is different from the results of the pilot study. The difference between self-enhancement and total bias, figured by the error bars, is significant for high anchor and high trust only.

- The panels in the middle row show that the bias from sensitivity S is positive or

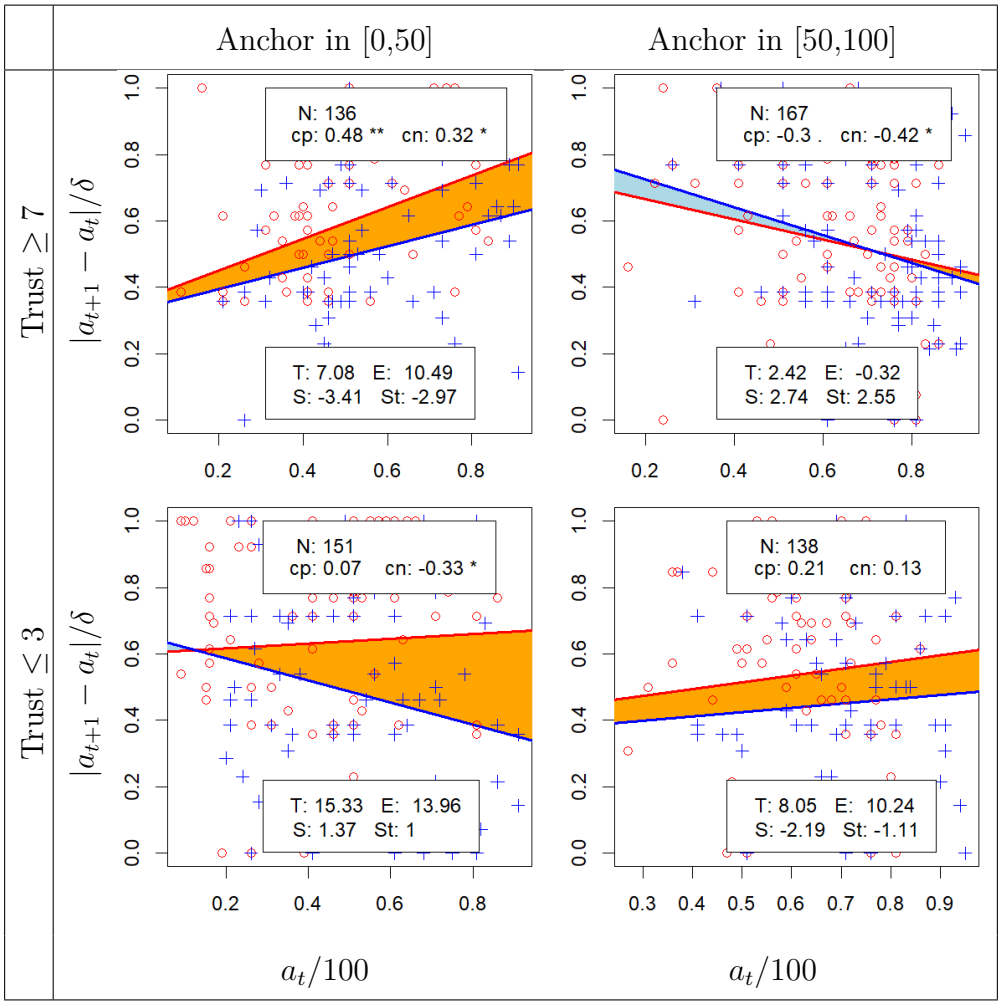

Figure 4. Examples of samples and regressions from pilot data. Red points are for positive feedbacks and blue crosses for negative feedbacks. c_p is the regression coefficient for positive feedbacks (red line). c_n is the regression coefficient for negative feedbacks (blue line). T is the total bias, E the self-enhancement bias, $S = T - E$ is the difference between the total and the self-enhancement biases, S_t the theoretical bias from varying sensitivity.

close to zero. Hence the main experiment does not confirm the negative S observed in the pilot study. However, like in the pilot study, S is the highest for high anchor and high trust. For the evaluation by rank, S is significant only for high anchor and high trust and it is close to 0 for low anchor and high trust. The average difference with the theoretical measure of the bias from sensitivity S_t (error bars) is 0.09 (standard deviation: 0.06).

- The panels of the bottom row show that for all samples the regression coefficient c_n is negative, with more negative values for score. For high anchor and high trust, c_n is significant for rank (*) and very significant for score (***). For most samples of low anchor (on the left), the coefficients c_p and c_n are of similar values and opposite signs. This explains why the bias from sensitivity is low, since $c_p + c_n$ is a factor in the expression of S.

Overall, focusing on the high trust case, these results show a significant bias from sensitivity to feedback for high anchor and a negligible one for low anchor (middle row of Figure 5).

Figures 6 and 7 show examples of samples and regressions for the combinations of low (≤ 40), high (≥ 60) anchors with low (≤ 2) and high (≥ 8) trust, for respectively rank and score evaluations. For both rank and score, coefficient c_n is significant only for high anchor and high trust (top right panels). For evaluations by rank, in all cases, the width of the orange area increases, hence self-enhancement increases, when a_t increases. Similarly, for evaluations by score, for high anchor, the width of the blue area decreases, hence the self-disparagement decreases, when a_t increases. For low anchor, when a_t increases, the blue area becomes orange at a given point, indicating a switch from self-disparagement to self-enhancement.

Effect of self-esteem. Figure 8 shows the different measures of bias and the regression coefficients, distinguishing rank and score like on figure 5, except that red error bars now show the difference of the same measure for participants of high self-esteem (i.e. higher than median value) only and the dark green error bars show the same measures for participants of low self-esteem (i.e. lower than median value) only. Moreover, there is a fourth row of panels devoted to regression coefficient c_p. The average size of the samples is 110.25 (standard deviation: 28.9) for low anchor and 334.25 (standard deviation: 81.0) for high anchor. Since each regression (for positive or negative feedback) is made on about half the samples, the results for low anchor are probably not very reliable and we mainly focus the analysis on the samples with high trust and high anchor.

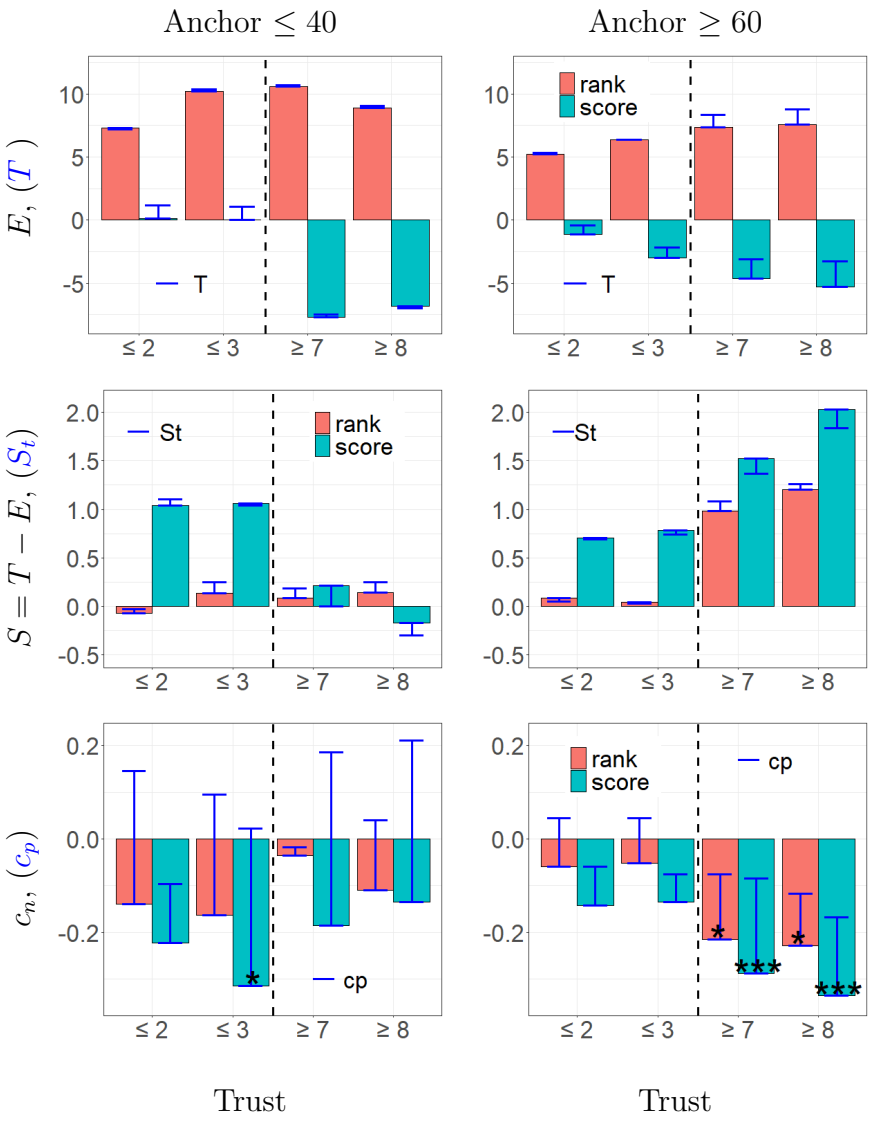

Figure 5. Main experiment data. Top row panels: Self-enhancement bias E, with error bars showing the difference to the total bias (T). Middle row panels: Bias from sensitivity $S = T - E$, with error bars showing the difference to the theoretical bias from sensitivity. Bottom row panels: Regression coefficients c_n, error bars showing the difference with c_p.

- In most cases (13 over 16), and in all cases for high anchor and high trust, the high self-esteem samples show a greater self-enhancement E than their low self-esteem counterparts.

- The bias from sensitivity S is higher for high than for low self-esteem, except for low anchor and low trust.

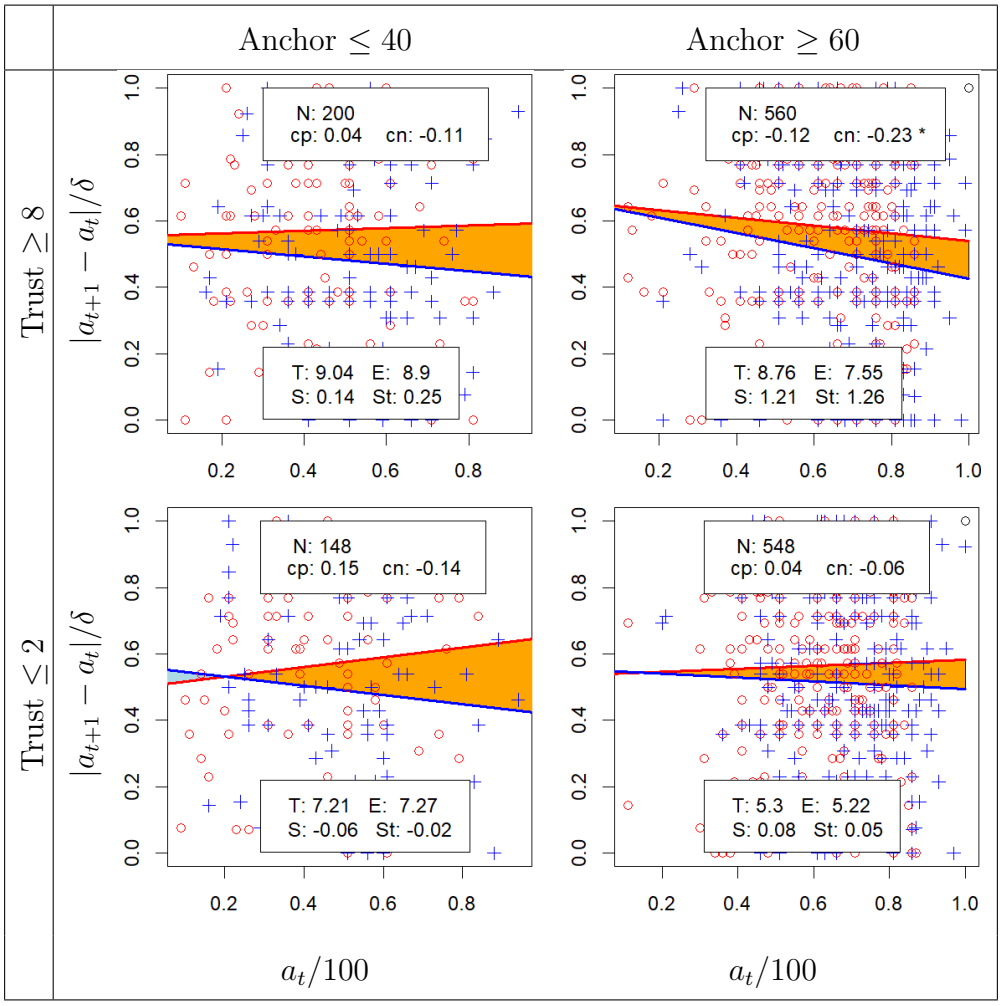

Figure 6. Evaluations by rank for main study data, examples of regressions and measures of different biases. The regression for negative feedbacks (blue line) significantly decreases only for high anchor and high trust (top right panel).

- The regression coefficients for positive feedbacks c_p are more negative for high than for low self-esteem, for evaluations by rank and by score.

- For evaluations by score, the regression coefficients for negative feedbacks c_n are more negative for high than for low self-esteem.

Figure 9 shows examples of samples and regressions for rank and score evaluations, high trust (≥ 8) and high anchor (≥ 60) and low and high self-esteem.

- For evaluations by rank and by score, the self-enhancement bias E is greater for high self-esteem, and the difference is larger for evaluations by score.

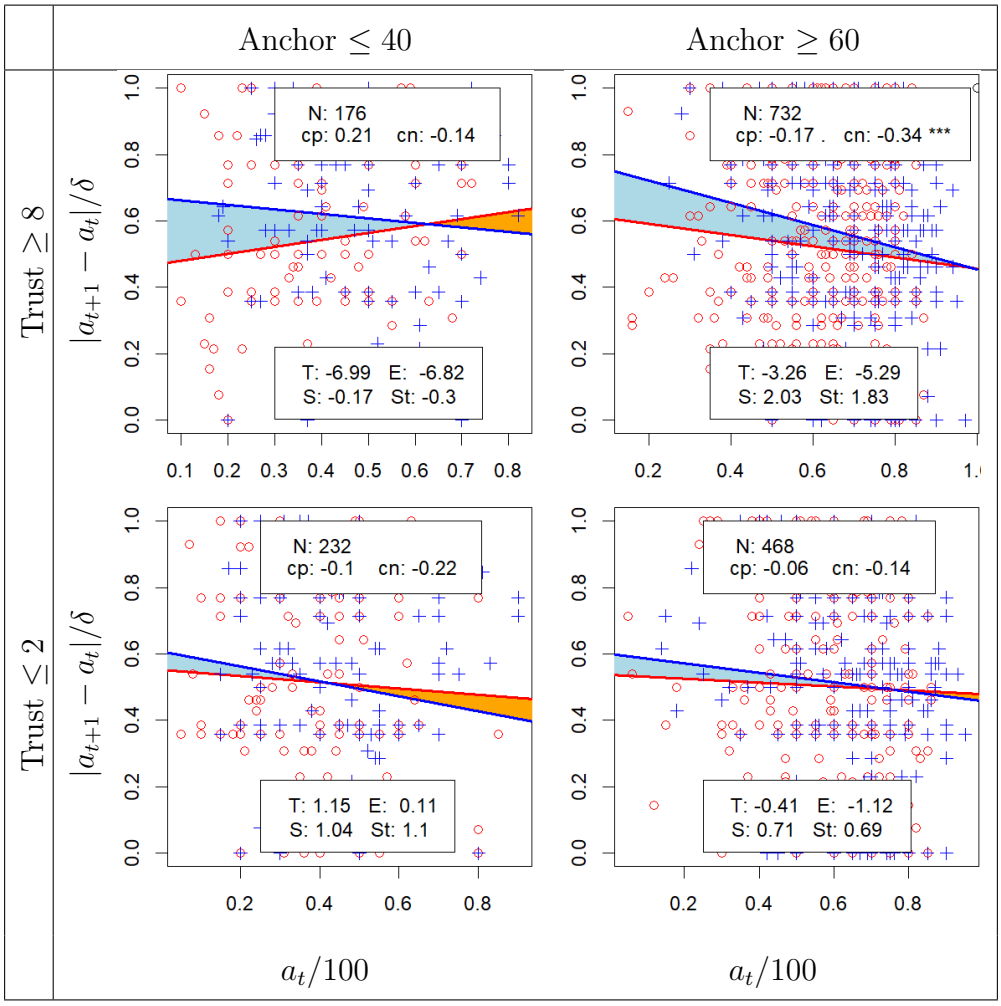

Figure 7. Evaluations by score for main study data, examples of regressions and measures of the different biases. The regression for negative feedbacks (blue line) significantly decreases only for high anchor and high trust (top right panel).

- For evaluations by rank, self-esteem does not seem to have an impact on the sensitivity bias, whereas for evaluations by score, the sensitivity bias is significantly higher for high self-esteem, because both coefficients c_p and c_n are more negative.

Discussion

In this section, we summarise the main results of the experiment, then we propose some technical comments about these results and finally, we discuss them in a broader perspective.

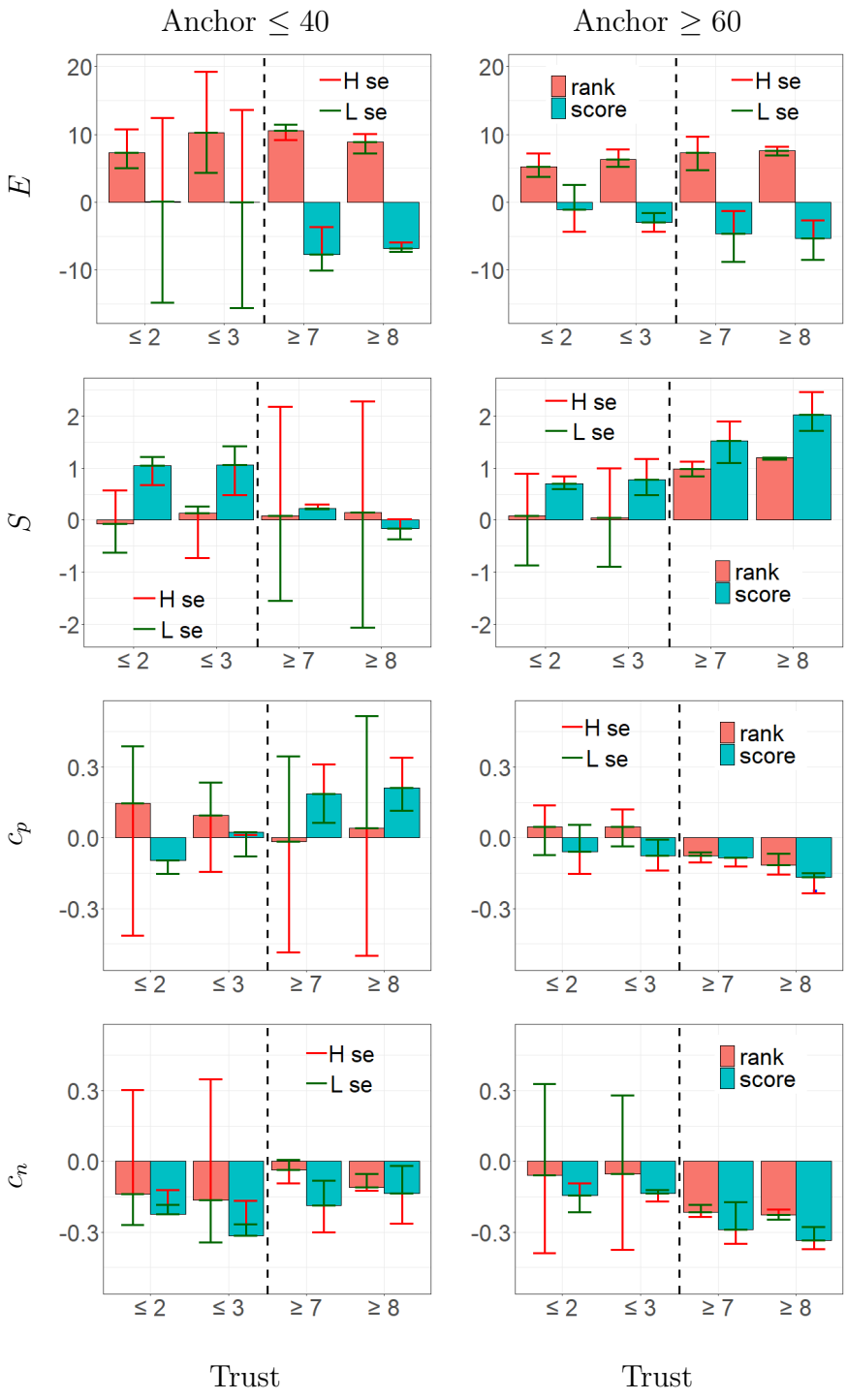

Figure 8. Effect of self-esteem. The measures of high self-esteem are represented by red error bars, the ones for low self-esteem are represented in by dark green error bars. More explanations in main text.

Summary and technical comments

The main purpose of this experimental work is to detect a positive bias coming from the decrease of the sensitivity to the feedbacks, as predicted by the theoretical

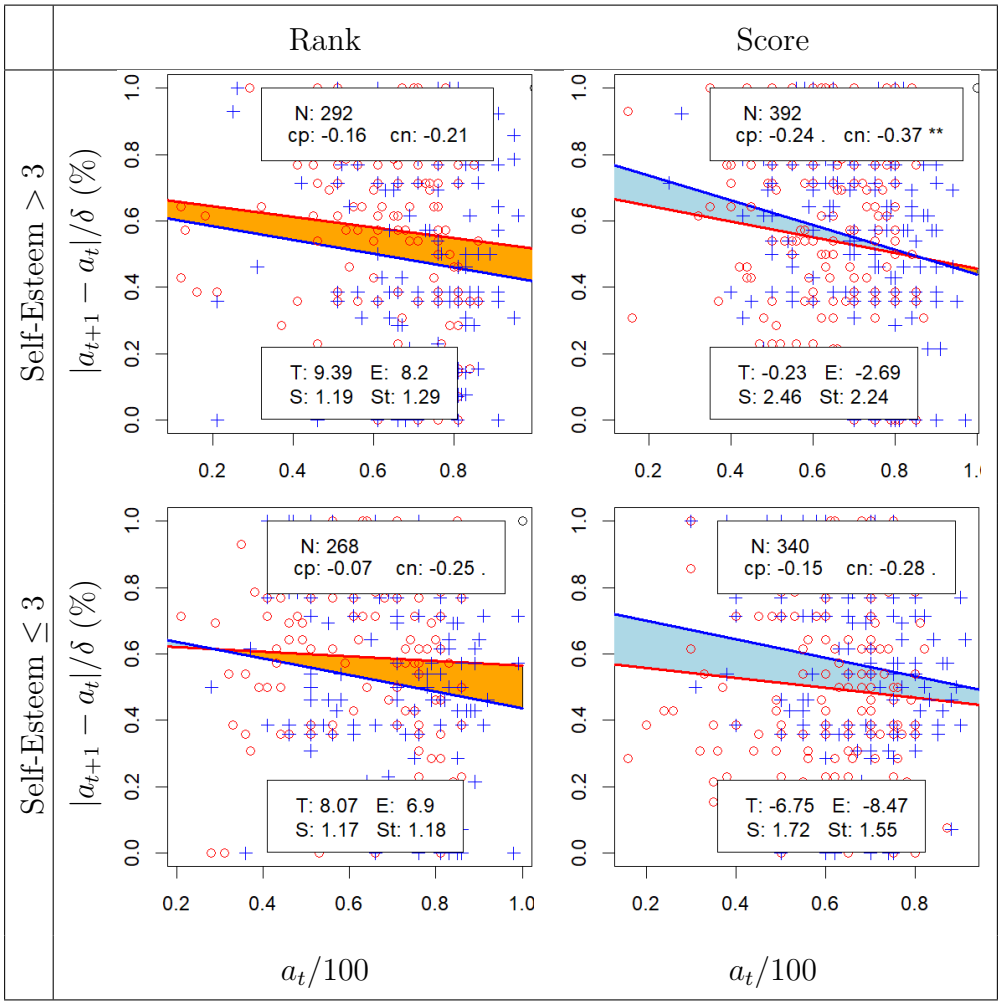

Figure 9. Effect of self-esteem for anchor ≥ 60 and trust ≥ 8. Top panels: self-esteem higher than median value. Bottom panels: self-esteem lower than median value.

model. In order to remain close to the context of the model, we mainly focus on the results for high trust (≥ 7 or ≥ 8) in the feedbacks, as the model assumes a complete trust in the feedbacks.

In both the pilot and the main experiments we found that the average sensitivity to both negative and positive feedbacks decreases when the self-evaluation increases for the samples of high anchor, this decrease being more significant for the negative feedbacks. In these conditions, we measure a bias from decreasing sensitivity ranging from 0.9 % to 2.9 % of feedback intensity δ. This measure is performed in two ways: one is the difference between the total and the self-enhancement biases ($S = T - E$) and the second (S_t) applies the mathematical formula coming from the theoretical

model. The average of $|S - S_t|$ on the tested samples without splitting between low and high self-esteem is 0.08 (standard deviation: 0.06) and when splitting between high and low self-esteem, it is 0.12 (standard deviation: 0.10). In our view, these results confirm the existence of a significant positive bias from decreasing sensitivity in the conditions of high trust and high anchor.

The reason why we detect a significant sensitivity decrease only in samples of high anchor is most probably that the sensitivity is a non-linear function of the self-evaluation and its slope is larger for high self-evaluation. Indeed, the values of the self-evaluations a_t in the triples (a_t, f_t, a_{t+1}) from a sample of high anchor are generally higher than the ones of a sample of low anchor. Therefore the interpretation of the results is that the bias from sensitivity to feedback is more likely to take place for participants of high self-evaluation.

Moreover, the bias from sensitivity to feedback is larger for score than for rank evaluations, because both the regression coefficients c_p (for positive feedbacks) and c_n (for negative feedbacks) are more negative for score than for rank evaluations. A possible explanation of this difference is that the change of self-evaluation is also a function of the relative feedback intensity $(f_t - a_t)/a_t$, and not of the absolute feedback intensity $f_t - a_t$ only, as assumed in the model. Indeed, when a_t increases, the relative intensity $\pm\delta/a_t$ of the feedbacks in our experiments becomes larger in absolute value for rank and smaller for score. This could therefore explain why the effect of the feedback tends to decrease more for score than for rank evaluations. This suggests to conduct experiments testing the impact of the relative intensity of the feedback.

The experiment shows additional features that were not expected from the theory:

- The self-enhancement bias is positive for the evaluation by rank and negative for the evaluation by score (hence it is rather a self-disparagement) in all trust and anchor conditions.

- Participants having a high self-esteem show a greater average self-enhancement (or a smaller average self-disparagement). For high anchor, and this difference between high and low self-esteem is stronger for score evaluations. The samples of

high self-esteem and high anchor tend also to show a high bias from sensitivity, particularly for score.

Broader theoretical implications

A positive bias on self-evaluation, without self-complacency? Sedikides and Strube (1997) define self-enhancement and self-assessment as follows:

- "self-enhancement is the motivation of people to elevate the positivity of their self-conceptions and to protect their self-concepts from negative information,
- self-assessment is the motivation of people to obtain a consensually accurate evaluation of the self."

Moreover, Sedikides and Strube (1997) stress that the positive bias on self-evaluation induced by self-enhancement is often considered useful because it can provide the will or general self-efficacy necessary to initiate novel action. As expressed by Cairns (1990): "Even if one is sick and anxious and poor, there should be reason to get up in the morning...Hence self-cognitions do not always have to be veridical in order to be functional". However, excessive self-overestimation can expose to severe negative consequences as shown in various domains such as health, education and the workplace (Dunning et al., 2004). Moreover, it can lead to self-complacency or bitterness when people become the only ones convinced of their high merit.

The motivation for self-assessment can be seen as moderating the excesses of self-enhancement. Indeed, self-assessment removes the protection against negative feedbacks in order to get an unbiased and accurate self-perception. There is thus a tension between both motivations as, in principle, an accurate self-assessment should remove the positive bias from self-enhancement as well as its advantages and drawbacks.

However, our work suggests that the self-assessment process, though removing protections against negative feedbacks, also generates a positive bias, when the sensitivity to feedbacks decreases as self-evaluation increases (i.e. according to our results, rather for people with an already above average self-evaluation). Moreover, this

positive bias seems adaptive to the context. Indeed, this bias takes place while there is an alternation of positive and negative feedbacks of similar intensity, therefore while the person's behaviour leads to evenly good and bad outcomes. However, if the change of behaviour due to an increase of self-evaluation leads to more frequent bad than good feedbacks, then the self-evaluation is immediately readjusted. This is less likely to be the case when self-enhancement tends to ignore negative feedbacks. Therefore, the positive bias from decreasing sensitivity pushes oneself forward as far as possible, in a cautious and adaptive way.

Moreover, the bias from decreasing sensitivity provides an incentive to accumulate different experiences. Indeed, this bias reinforces confidence while the experiences have at least 50 percent chances to succeed, hence it rewards active and risky behaviours. There is a substantial difference with self-enhancement that is likely to privilege caution in order to avoid negative feedbacks.

Effect of evaluation scale and self-esteem. The negative self-enhancement (or rather self-disparagement) observed with the evaluation by score was a surprise. We can only formulate a preliminary hypothesis that could explain it. The score could be perceived as a possessed quantity, as much as the evaluation of an ability. Then, a negative feedback would thus be perceived as the loss of some possessions as well as a set back in status. The higher reaction to negative feedbacks could then be related to a general loss aversion (Kahneman & Tversky, 1979) or higher sensitivity to repeated negative events (Baumeister, Finkenauer, & Vohs, 2001). By contrast, the evaluation by rank seems to be more exclusively related to a perceived status, triggering self-enhancement, as expected from the literature.

This impact of the scale of evaluation could be related to the experiment reported by Cherry and Ellis (2005), suggesting that student performance is significantly improved when facing a grading system based on student ranking (norm-reference grading) rather than performance standards (criterion-reference grading). In this experiment, the grades are letters from A to F, which can be nuanced with plus or minus. In the grading based on student ranking, the different grades are directly defined

by the ranking while the other grading is only based on performance standards. From our experiment, it seems possible that the grading systems generate significantly different self-enhancements or self-disparagements, which could influence the performance of the students. In particular, strong levels of self-disparagement which, extrapolating from our results, could be expected with the grades based on performance, could discourage students and be detrimental to their performance. Of course, this does not exclude the likely influence of other factors mentioned by Cherry and Ellis (2005).

Further, our results (see Figure 8) showing that people with a high self-esteem or a high self-evaluation tend to show a greater self-enhancement (for the evaluation by rank) or a smaller self-disparagement (for the evaluation by score) are corroborated by other studies (Bosson, Brown, Zeigler-Hill, & Swann, 2010; Kobayashi & Brown, 2003). Moreover, self-determination and terror management theories propose potential explanations of this relation between self-esteem and self-enhancement (Hohman & Brown, 2020).

Finally, it seems noticeable that our main result, the existence of a bias from sensitivity to feedback, originates in theoretical agent simulations. This bias was indeed identified because its effects were easily observable in long lasting simulations, involving millions of virtual interactions. We could then observe its much smaller effect on short simulations, that we had initially overlooked. Similarly, it seems almost impossible to observe this bias in real life without looking for it in a specific experiment. This is a case, common in physics but not so much in social sciences, of an initially purely theoretical concept whose existence is confirmed experimentally.

Acknowledgements

This work has been partly supported by the ORA project ToRealSim. We are grateful to Sylvie Huet for her help at early stages of the research.

References

Baumeister, R., Finkenauer, C., & Vohs, K. (2001). Bad is stonger than good. *Review of General Psychology*, *5*(4), 323-370.

Bosson, J., Brown, R., Zeigler-Hill, V., & Swann, W. (2010). Self-enhancement tendencies among people with high explicit self-esteem: The moderating role of implicit self-esteem. *Self and Identity*, *2*(3), 169-187.

Buchler, R., Griffin, D., & Ross, M. (1991). Depression, realism, and the over-confidence effect: Are the sadder wiser when predicting future actions and events? *Journal of Personality and Social Psychology*, *67*, 366-361.

Cairns, R. B. (1990). Theories of the evolution of knowing: Thel c. schneirla conference series. In G. Greenberg & E. Tobach (Eds.), (p. 69-86). Hillsdale, NJ: Erlbaum.

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*, 23-43.

Cherry, T., & Ellis, L. (2005). Does rank-order grading improve student performance? evidence from a classroom experiment. *International Review of Economics Education*, *4*(1), 9-19.

Deffuant, G., Bertazzi, I., & Huet, S. (2018). The dark side of gossips: hints from a simple opinion dynamics model. *Advances in Complex Systems.*, *21*.

Deffuant, G., Carletti, T., & Huet, S. (2013). The leviathan model: Absolute dominance, generalised distrust and other patterns emerging from combining vanity with opinion propagation. *Journal of Artificial Societies and Social Simulation*, *16*(23).

Dunning, D., Heath, C., & Suls, J. (2004). Flawed self-assessment. implications for health, education and the workplace. *Psychological Science in the Public Interest*, *21*, 69-106.

Dunning, D., & Story, A. (1991). Exploring the planning fallacy: Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *61*, 521-532.

Epley, N., & Dunning, D. (2000). Feeling holier than thou: Are self-serving assessments

produced by errors in self or social prediction? *Journal of Personality and Social Psychology*, *79*, 861-875.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552-564.

Flache, A., Maes, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4).

Griffin, D., Dunning, D., & Ross, I. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology*, *59*, 1128-1139.

Hohman, Z., & Brown, J. (2020). Self-esteem and self-enhancement. *Psychology*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Ecnometrica*, *47*(2), 263-291.

Kobayashi, C., & Brown, J. (2003). Self-esteem and self-enhancement in japan and america. *Journal of Cross-Cultural Psychology*, *34*(3).

Moreland, R. L., & Sweeney, P. D. (1984). Self-expectancies and relations to evaluations of personal performance. *Personality*, *52*, 156-176.

Sedikides, C., & Strube, M. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances In Expermiental Social Psychology*, *9*, 209-269.

Vallières, E., & Vallerand, R. (1990). Traductuion et validation canadienne-française de l'échelle de l'estime de soi de rosenberg. *International Journal of Psychology*, *25*(2), 305-316.

Vallone, R., Griffin, D., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology*, *58*, 582-592.