Labour Economics 76 (2022) 102166

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

Image concerns in ex-ante self-assessments–Gender differences and behavioral consequences☆

Simone Haeckl[a,b,*]

[a] University of Stavanger, Business School, Stavanger, Norway
[b] WU Vienna, Department of Economics, Vienna, Austria

ARTICLE INFO

JEL classification:
C91
D91
J16

Keywords:
Bias
Gender
Overconfidence
Real-effort experiment
Self-assessment

ABSTRACT

While differences in confidence have been identified as a driver behind gender gaps in the labor market, less is known about the moderators of these differences. This paper uses a laboratory experiment to investigate how the image concerns related to the self-assessment affect male and female confidence. Subjects assess their relative performance prior to a real-effort task and can subsequently adjust their efforts. I find that women increase their self-assessment when it is made public, but only if the actual placement remains private. There are no effects for men. I also investigate whether subjects who ex ante overstate their ability provide more effort. I find no evidence of such a motivational effect.

1. Introduction and literature

Gender differences in confidence are well-documented and have been used to explain differences in labor or career-related decisions like the choice of major at university or the willingness to enter a competitive work environment resulting in the gender wage gap (Buser et al., 2017; Croson and Gneezy, 2009; Gneezy et al., 2009; 2003; Niederle and Vesterlund, 2007; 2008; van Veldhuizen, 2018). Gender differences in confidence may also affect decisions on the demand side of the market. For example, if employers do not sufficiently account for gender differences in confidence when making their hiring decisions, they may discriminate against women whose self-assessments are less confident. Such discrimination would also be costly if it prevents employers from selecting the most productive candidates (Reuben et al., 2014).

While gender differences in confidence and their consequences are well studied, less is known about the moderators of these differences.[1] This is surprising as previous literature suggests that the level of observed overconfidence differs by numerous factors, such as the difficulty of the task (Moore and Healy, 2008) or the image concerns related to the self-assessment (Ewers and Zimmermann, 2015). Moreover, it has been found that gender differences in confidence seem to increase if the task is considered to be stereotypically male (Bordalo et al., 2019; Coffman, 2014), or if overconfidence can be observed by others (Ludwig et al., 2017).

In this paper, I study whether the mode of self-assessment, either in public or private, affects men and women differently. I run a real-effort experiment in which subjects ex-ante self-assess their performance relative to others (i.e., their placement) in a familiar and simple math task. Following previous literature, I modify the image concerns related to the self-assessments by having treatments that differ in whether or not self-assessments are public and verifiable by others (Ewers and Zimmermann, 2015). In contrast to previous experiments, the self-assessment is elicited prior to the relevant real-effort task and I implement a setting in which subjects can work longer to increase output and live up to their assessment.[2]

Giving subjects the possibility to increase effort to live up to their self-assessments is an important extension of previous literature for three reasons. First, it reflects practice in the professional world. Con-

☆ I gratefully acknowledge support by the Vienna University of Economics and Business, both financially and for providing access to the experimental infrastructure at WULABS. I appreciate helpful comments on the paper from Ben Greiner, Gergely Hajdu, Melis Kartal, Mari Rege, Rupert Sausgruber, Jean-Robert Tyran, and Roel van Veldhuizen.

* Corresponding author.
E-mail address: simone.haeckl-schermer@uis.no

[1] A notable exception is Exley and Kessler (2019) who run a series of experiments to identify moderators of gender differences in self-promotion.

[3] The task is characterized by avoiding corner solutions in effort-provision by introducing monetary opportunity cost of time (see Section 2 for details).

https://doi.org/10.1016/j.labeco.2022.102166
Received 15 May 2021; Received in revised form 24 February 2022; Accepted 31 March 2022
Available online 11 April 2022

sider for example a job interview. While the employer might be able to infer your performance in previous jobs based on references, she would actually like to know, and might also ask you, how well you will be doing on the tasks in your new job. These tasks are probably similar to the tasks you have performed in your previous job but you are able to affect your performance in the upcoming tasks based on the effort you are going to provide. This feature is captured by the elicitation of ex-ante self-assessments on a familiar task in which performance depends on ability and effort. Second, the change in timing might decrease observed gender differences in confidence if they are caused by women who are afraid of receiving negative feedback (Croson and Gneezy, 2009; Niederle and Vesterlund, 2007) because negative feedback can be avoided by increasing effort. Third, comparing output between treatments with ex-ante and ex-post self-assessments allows me to identify whether there is a motivational effect of an (overconfident) ex-ante self-assessment. Together, the setup contributes to the literature on the gender gap in career success by identifying moderators of gender differences in observed confidence. This is a first step toward developing measures to mitigate the distortionary effects of overconfidence in the labor market.

The main findings of my paper are as follows. Consistent with the previous literature, I observe that many subjects tend to overplace themselves, i.e., they overestimate their placement. Overall, I qualitatively replicate the results of Ewers and Zimmermann (2015), that self-assessments are higher when they are made public rather than private and lower when there is public feedback, however, this pattern is overwhelmingly driven by women. The finding that women state higher self-assessments when they are made public indicates that overplacement is driven by a preference to signal ability to others rather than by biased beliefs about own performance.[3] In addition, it suggest that social approval might be more important for women than for men. I also find that self-assessments are significantly higher when elicited prior to the real-effort task as compared to ex post. Again this effect is only driven by women. Finally, I investigate whether subjects who are ex ante too optimistic regarding their relative performance try to live up to their self-assessments and work harder in comparison to treatments in which subjects assess themselves ex post. I do not find any evidence for such behavior.

Relation to the literature. Most closely related to my paper is Ewers and Zimmermann (2015) who show that self-assessments change if they are made public and if public feedback is provided. These authors find that making self-assessments public induces a positive shift of self-assessments; they also find that this effect only occurs when the actual performance remains private, i.e., there is no effect when feedback is public. This result suggests that subjects overstate their placement because of image concerns. Schwardmann and van der Weele (2019) go one step further and suggest that subjects might deceive themselves, i.e., make themselves believe that they are of high ability, to be able to convince others. In a lab experiment, they show that informing subjects that they will have to convince an employer of their ability increases overconfidence in a private self-assessment. While there is no strategic component to the self-assessment in my experiment, i.e., there are no employers, subjects might still want to inflate their beliefs to convince observers of their ability. Another close match to this paper is Exley and Kessler (2019). In a series of experiments, the authors identify different moderators of gender differences in self-promotion. In contrast to this paper and previous literature, they ask subjects to evaluate their performance on a subjective scale (using adjectives) rather than on an objective scale (using, e.g., output quantiles). They find that gender differences in self-promotion are very persistent and cannot be explained by differences in objective performance beliefs.

I add to previous research by explicitly testing for gender differences in image concerns and by investigating how timing affects self-assessments. Precisely, I investigate whether self-assessments change if subjects can affect their placement by comparing self-assessments in a treatment in which self-assessments are elicited before the real-effort task to a treatment in which self-assessments are elicited after the real-effort task.

The focus on gender differences in image concerns is important, because it may affect employment decisions. This holds true particularly if differences in self-assessments increase with the level of observability. Ludwig et al. (2017) show in a principal-agent setting that women's self-assessments decrease when principals can observe the accuracy of their self-assessments, while men do not react. The authors provide evidence that women decrease their self-assessment significantly more than men because they are more susceptible to the feeling of shame. In comparison to their study, my results are obtained without a principal being affected by the self-assessment and, therefore, in the absence of social responsibility toward other subjects. Thereby my design allows me to test for gender differences in the feeling of shame from overstating one's ability without social preferences potentially confounding the effect. Another important reason to consider gender differences in image concerns and the resulting differences in confidence is provided by Niederle and Vesterlund (2007). The authors show that overconfidence plays an important role in understanding gender differences in the willingness to enter competition using lab experiments.[4] Understanding the roots of differences in the willingness to compete is crucial as they are an important determinant of labor market success (Buser et al., 2014). Buser et al. (2021) investigate whether gender differences in the willingness to compete (rather than confidence) change if the decision to enter a competition is made public and if there is public feedback on the outcome of the competition. They find suggestive evidence that men are more likely to enter competition when the decision is made public. However, in general, the effects of observability are negligible in their experiment. This result suggests that observability does not affect the size of gender differences in competitiveness. I add to their paper by investigating gender differences in self-assessments rather than competitiveness.

In addition, I investigate whether subjects are willing to exert additional effort to live up to an overly optimistic ex-ante self-assessment, i.e., overplacement, adding to the discussion of a motivational value of overconfidence (Bénabou and Tirole, 2002). Typically, the motivational value of overconfidence is based on the idea that overconfident individuals overestimate their gains to effort. Chen and Schildberg-Hörisch (2019) test for the motivational value of overconfidence in a real-effort experiment. They find that subjects' effort is indeed associated with overconfidence. In particular they provide evidence that subjects who spend more time on the real-effort phase do so because they overestimated their return to effort. However, they also find that providing feedback on the accuracy of the self-assessment reduces overconfidence and effort provision. Therefore, the findings of Chen and Schildberg-Hörisch (2019) are consistent with subjects having unconscious, erroneous self-assessments. An overestimation of the returns to effort is ruled out by design in my experiment as I continuously provide subjects with feedback on their output. In contrast, I investigate whether conscious overplacement results in an increase in effort.

A motivational effect of an ex-ante self-assessment is plausible if subjects wish to avoid feelings of shame from overly optimistic self-assessments, in particular, when feedback is public (Ludwig et al., 2017). It can also be rationalized if the ex-ante self-assessment sets a reference point, as discussed by Köszegi and Rabin (2006, 2007) and

[3] The terminology is based on Moore and Healy (2008) who distinguish between three types of overconfidence: overestimation, overplacement, and overprecision. As subjects assess their ability relative to others in my design, I use the term overplacement.

[4] van Veldhuizen (2018) shows that overconfidence might be even more important to explain these differences than previously assumed. In a recent paper Brandts et al. (2020) also propose gender differences in status-ranking aversion as a source for differences in competitiveness.

Table 1
Experimental schedule.

Phase	Description: Private treatment	Treatment variations
IQ/RA	20 min cognitive ability task (piece rate) and elicitation of risk aversion (lottery)	
Learn	4 min real-effort task (no monetary incentives)	
Ph0	5 min real-effort task (piece rate)	
Ph1	20 min real-effort task (piece rate + opportunity cost of time)	
Self- assessment	Private self-assessment on relative performance in phase 2 (no monetary incentives)	+ public self-assessment in **Audience** and **Feedback**
Ph2 (piece rate + opportunity cost of time)	20 min real-effort task	
Feedback (no monetary incentives)	Private feedback on the accuracy of the self-assessment	
Feedback	+ public feedback in	

Notes: The table shows the experimental schedule for ex-ante treatments, i.e., treatments in which the self-assessment takes place prior to the real-effort task. In addition, there is an ex-post treatment in which the self-assessment happens after Ph2.

Abeler et al. (2011) because not living up to the self-assessment will be perceived as a loss. In this way, reference points are similar to endogenously set performance goals, which have been shown to increase effort provision if set ambitiously (Locke and Latham, 1990). In a series of field experiments, Goerg and Kube (2012) investigate, among other things, the effect of non-binding and non-incentivized goals and confirm a performance-increasing effect. In a similar vein, Koch and Nafziger (2011, 2020) provide a theoretical model as well as experimental evidence showing that non-binding self-set goals can increase effort provision and may be used as a commitment device for present biased individuals. Non-incentivized, endogenous goal-setting has also been used to motivate students' performance, e.g., van Lent and Souverijn (2017) or Clark et al. (2017). Both studies find positive effects on performance in final exams. However, the effects differ both by size and statistical significance, indicating that the effectiveness of goal-setting may depend on the specific setting.

2. Experimental design

I use a multi-phase experiment (see Table 1). At the beginning of the experiment, I measure proxies for subjects' risk aversion and cognitive ability, both of which could explain differences in self-assessments (see IQ/RA below).[5] Afterwards, subjects work on a number of real-effort phases (Learn, Ph0, Ph1, and Ph2), provide a self-assessment of their relative performance, and receive feedback on the accuracy of their self-assessment. While the real-effort phases are constant across treatments, the timing (ex ante or ex post) and type of self-assessment as well as the way feedback is provided (private or public) vary by treatment. I now explain the general design, before I discuss the treatment variations. Incentives are presented as points with 15 points being equal to 1 Euro. The instructions are provided in Appendix F.

Phase IQ/RA. After arriving at the lab, subjects take a short, incentivized version of Raven's Advanced Progressive Matrices test. For each correctly solved matrix subjects receive 2 points. I use performance on the matrices as a proxy for cognitive ability. Subjects are paid for correct answers in the cognitive ability task to reduce the confounding effect of motivation in measuring cognitive ability (Borghans et al., 2009). Next, I elicit risk preferences using a lottery-based multiple price list with constant probabilities (see, for example, Drichoutis and Lusk, 2016). At the end of the experiment, one of the subjects' choices in the price list is randomly picked for payment. To reduce potential spill-over effects of the measures elicited in this phase on the main experimental measures subjects neither receive any feedback on their performance on the test nor learn their payoff from this phase until the end of the experiment. Also, phase IQ/RA is identical across treatments.

Real-effort phases. Subjects work on a simple but tedious task consisting of adding single-digit numbers (cross sums).[6] On each screen, subjects have to calculate three tasks, and after every fifth screen, an additional digit is added to the tasks, which makes the calculations more difficult. If the answer to one of the tasks is not correct, subjects receive a hint as to which of the tasks is wrong. They can only proceed to the next screen after solving all three tasks correctly. There are four real-effort phases (Learn, Ph0, Ph1, and Ph2), all of which will be explained in detail below.

- Learn and Ph0: Phase Learn serves to familiarize subjects with the task. This phase lasts for four minutes and subjects do not yet receive payment in this phase. In Ph0, subjects work on the same task for five more minutes, now being paid a piece rate of 0.7 points per correct task, or as there are three tasks per screen, 2.1 points per screen. I will use subjects' output in this phase as a proxy for ability.
- Ph1 and Ph2: Phases 1 and 2 each last for 20 minutes. The task and the piece rate are equivalent to Ph0. However, incentives differ in that there are monetary opportunity costs of time as subjects can stop working.[7] Once they have stopped working, subjects do not calculate anymore tasks but remain seated in front of their screens. For the remaining 20 minutes, i.e., until the end of the phase, they receive 1 point every 15 seconds. As discussed by Haeckl et al. (2018), the combination of an increasingly time-consuming task with a paid outside option enables me to estimate a production function for each individual based on their ability and to predict a benchmark for the money-maximizing switching point at the individual level. A money-maximizing subject should stop working on the task as soon as it takes more than 31.5 seconds (2.1 points per screen * 15) to complete a screen. The parameters are chosen based on data from a previous experiment to ensure that subjects have an interior money-maximizing switching point.[8] An interior switching point is necessary to allow subjects to increase effort as a response to their self-assessment. I provide subjects with feedback on how long they took to solve the previous screen. In this way, I ensure that the decision to deviate from money-maximizing behavior is deliberate and is not caused, for example, by a biased perception of time.[9] I use deviations

[5] Cognitive ability and risk aversion are elicited at the beginning of the experiment as the main part of the experiment is cognitively demanding and the effort subjects provide in the main part of the experiment might vary by treatment which could have biased the measure of cognitive ability.

[6] See also Haeckl et al. (2018) for an application of the task in a different context in a previous experiment with a different subject pool.

[7] Several studies suggest to use an incentivized alternative task to introduce monetary opportunity cost of work in real-effort experiments (e.g., Berger et al., 2013; Blumkin et al., 2010; Eckartz, 2014; Erkal et al., 2018; Gächter et al., 2016; Hayashi et al., 2013; Mohnen et al., 2008; Weber and Schram, 2017) and by that to allow subjects to react to changes in incentives.

[8] Also in the current experiment no subject should work the entire time if they want to maximize their earnings. On average they should switch after 552 (out of 1200) seconds.

from this money-maximizing benchmark to investigate the motivational effect of an overconfident self-assessment.

Self-assessment. Subjects' beliefs on their placement in the real-effort task are elicited in private before treatments are implemented. In the ex-ante treatments shown in Table 1, subjects assess their placement in the upcoming real-effort task before phase 2 starts. In the ex-post treatment, subjects do the assessment after phase 2. In line with previous literature, subjects assess themselves relative to an unknown but similar group (see e.g., Ewers and Zimmermann, 2015). This reference group consists of participants in an earlier experiment using the same real-effort task and is introduced to avoid that subjects' beliefs about their placement depend on their beliefs on other subjects' reactions to the ex-ante self-assessment which likely vary by treatment. Subjects have to indicate whether they think that they are better or worse than the average subject at the task. Subjects are told that the correctness of their self-assessment will be evaluated based on the number of cross-sums they have solved in phase 2 to align the understanding of what the self-assessment should be about.[10] I also elicit the belief distribution by asking subjects how likely they think it is that they will fall into each of four output quartiles, ranging from the lowest 25% to the highest 25%. To make the elicitation as intuitive as possible, I show subjects a slider for each quartile and ask them to distribute the 100 percentage-points over these four quartiles (see instructions in Appendix F).[11] The correctness of the self-assessments is not incentivized.

A caveat of this design choice is that it might increase the scope for inflated beliefs (Charness et al., 2021). However, using an incentive-compatible belief-elicitation method might introduce other biases (Benoît et al., 0000), increases the complexity of the experiment and does not necessarily increase the reliability of the elicitation (Charness et al., 2021; Schlag and Tremewan, 2021; Trautmann and van de Kuilen, 2015). As the focus of the experiment is to identify gender-specific treatment differences in the self-assessments (keeping the elicitation method constant), rather than to measure absolute levels of beliefs, I abstained from incentivizing the belief elicitation. In addition, informing subjects in the treatment with an ex-post self-assessment about the upcoming self-assessment could have affected their effort and thereby confounded my results and not informing subjects about an incentivized ex-post self-assessment could be perceived as deception.[12]

Feedback. In all treatments, after phase 2 subjects receive private feedback on their relative output quartile as well as on whether their output is higher or lower than the average subject's output of a similar group.

Treatments. I implement four treatments: *Private, Audience, Feedback*, and *Ex post* (see Table A.1 in Appendix A for an overview). Similar to Ewers and Zimmermann (2015), the treatments differ in whether the self-assessments and the feedback are observable by others or not. Subjects are informed about the mode (private or public) of the self-assessment and the feedback before making their private self-assessments. The *Private* treatment is exactly as described in Table 1. Treatment *Audience* is equivalent to the *Private* treatment except that after finishing the private self-assessment, subjects must stand up and publicly announce whether they believe that they are better or worse than average. The order in which subjects publicly announce their self-assessments is randomly determined. I ask subjects to first privately enter their self-assessment into the computer before publicly announcing their self-assessment to avoid order effects. In treatment *Feedback*, subjects not only announce their self-assessment publicly but they also receive public feedback. Regarding the latter, subjects have to stand up again after the real-effort task and the experimenter publicly announces whether they were better or worse than the average subject of the reference group and whether their self-assessment was correct.[13] To investigate whether timing affects the self-assessment, and to identify the motivational effect of overconfident ex-ante self-assessments, I also run an ex-post version of the *Audience* treatment in which the self-assessment is elicited after phase 2.

Procedure. Experiments were conducted between 2018 and 2019 in the laboratory of the Vienna University of Economics and Business (WU) using z-Tree (Fischbacher, 2007). A total of 345 students were recruited using ORSEE (Greiner, 2015). These 345 students were grouped into 12 sessions based on their availability and each session was randomly assigned to one of the four treatments with three sessions per treatment. Precisely, 89 students participated in Private, 84 in Audience, 82 in Feedback, and 90 in Ex post. Three subjects are excluded from the analysis because of technical problems during the experiment reducing the sample to 342 students. The language used was English. The experiment took 90 minutes, and subjects earned an average of € 25. To achieve a gender-balanced sample, an equal share of men and women were invited to each session. Gender balancing worked well over all treatments, as approximately half of the subjects were female (52.05%). Most of the subjects were studying economics, business, or business law (73.78%) and were Austrian or German citizens (66.13%).

3. Model and hypotheses

To get an intuition into how the different treatments might affect self-assessments and effort provision, consider a very stylized and simple model. Consider first behavior without a self-assessment. In the simplest case (ignoring all non-monetary outcomes), the utility of individual i depends only on her earnings $\pi_i(Q_i)$ with Q_i representing the number of completed tasks. In this case, i's utility is simply

$$U_i(Q_i) = \pi_i(Q_i). \tag{1}$$

I assume that $\pi_i(Q_i)$ is a concave function and that there is an optimal number of completed tasks Q_i^* that maximizes the individual's earnings.[14] An individual has a belief about her relative ability $A_i(Q_i^*, \alpha_i Q_{-i})$. This belief depends on Q_i^* and her belief about the average output of a similar group. While the actual output of this group Q_{-i} is determined exogenously, the belief about the average output might be biased by a factor of α_i. This factor determines the individual's level of confidence, with $\alpha_i \geq 0$. If $\alpha_i < 1$, the individual underestimates the average ability and is overconfident, if $\alpha_i > 1$ the individual overestimates the average ability and is underconfident, and if $\alpha_i = 1$ the individual has an accurate assessment.[15] I assume that the belief about one's relative ability, $A_i \in \{0, 1\}$, is 1 if individual i believes she is better than average, i.e., if $Q_i^* > \alpha_i Q_{-i}$, and 0 otherwise.

[9] An overconfident subject could underestimate the time it takes her to complete a screen and, thereby, overestimate her gains to effort. By providing continuous feedback on the time per screen, I reduce the scope of erroneous beliefs about gains to effort.

[10] Still, I cannot rule out that subjects had a more general concept of ability in mind when making the self-assessment.

[11] This way of eliciting the distribution of beliefs is well established in the literature (see Eil and Rao, 2014; Grossman and Owens, 2012; Moore and Healy, 2008).

[12] In contrast to previous experiments with ex-post self-assessments, the real-effort task is specifically designed to allow subjects to affect their ranking by increasing effort. This design choice makes the information about an incentivized self-assessment potentially behaviorally relevant and thereby withholding this information deceptive (see e.g., Krawczyk, 2019).

[13] For simplicity, the names of the treatments are consistent with Ewers and Zimmermann (2015).

[14] In the experiment, $\pi_i(Q_i)$ is the sum of earnings from working on the real-effort tasks and the money earned as a flat rate while sitting idle. While the earnings from working on the real-effort task increase in Q_i, the earnings from being idle decrease in Q_i as the time left for sitting idle decreases the more time a subject spent working on the task. See also Haeckl et al. (2018) for a formalization of the optimization problem and an example of how money-maximizing behavior can be estimated.

[15] For simplicity, I assume that the individual is certain that her belief is true and abstain from introducing uncertainty to the model.

In the first stage of the experiment, subjects have to state whether they believe that they are better or worse than the average subject of an exogenous comparison group. For simplicity, let us assume that the self-assessment $SA_i \in \{0,1\}$ with $SA_i = 1$ if the individual states that she is better than average and 0 otherwise. As discussed, for example, in Burks et al. (2013), an individual can be prone to image concerns ($s_{iTMT} \geq 0$). How much an individual cares about her image depends on the treatment (TMT), as will be discussed below when I derive my hypotheses. The image utility $s_{iTMT}SA_i$ reduces to s_{iTMT} if the individual states that she is better than average, i.e., $SA_i = 1$, and there is no image utility, if the individual states a low self-assessment, i.e., $SA_i = 0$.[16] If an individual believes that she has a high ability, $A_i = 1$, and maximizes her earnings, i.e., $Q_i = Q_i^*$, her utility from stating a high ability, $SA_i = 1$ is

$$U_{i_{A_i=1,SA_i=1}}(Q_i^*) = \pi_i(Q_i^*) + s_{iTMT}. \tag{2}$$

If she states ($SA_i = -1$), her utility is

$$U_{i_{A_i=1,SA_i=-1}}(Q_i^*) = \pi_i(Q_i^*). \tag{3}$$

As a result, an individual who believes that she is better than average should state a high ability if $s_{iTMT} > 0$.[17] That is, individuals who believe they are better than average will always state their beliefs.

In contrast, individuals who believe they are worse than average might decide to misreport their beliefs. That is, if an individual believes that she is worse than the average participant ($A_i = 0$), she might still decide to state a high self-assessment to get the positive image utility. However, as subjects receive feedback at the end of the experiment, the positive utility from signaling high ability in the self-assessment can be offset by the feeling of shame λ_{iTMT}, with $\lambda_{iTMT} \geq 0$:

$$U_{i_{A_i=0,SA_i=1}}(Q_i^*) = \pi_i(Q_i^*) + s_{iTMT} - \lambda_{iTMT}. \tag{4}$$

The individual expects to feel shame if her expected ranking in the real-effort task $R_i(Q_i, \alpha_i Q_{-i})$, is worse than her self-assessment. I assume that $R_i \in \{0,1\}$, where $R_i = 1$ if individual i believes she will complete more tasks than average, i.e., if $Q_i > \alpha_i Q_{-i}$, and 0 otherwise. The dis-utility from shame is $\lambda_{iTMT}(-1 + R_i)$. If the individual maximizes her earnings the expected ranking is equal to the ability belief, i.e., $R_i = A_i$. If the individual has a high ability ($A_i = 1$) and produces Q_i^*, there is no shame from stating a high ability as $R_i = 1$. If she has a low ability ($A_i = 0$) and produces Q_i^*, $R_i = 0$, her loss in utility due to the feeling of shame after stating a high ability is λ_{iTMT}. It follows that an individual with low ability who produces the money-maximizing output Q_i^*, has a utility from stating a high self-assessment as shown in Eq. (4).

However, as the real-effort phase happens after the self-assessment, the individual could also adapt her effort. I assume this is a two-step decision process, and the individual decides whether to state a high or a low ability using backward induction. Precisely, when deciding which ability she wants to state, the individual knows that she can increase output to what she believes is the average output plus one additional task, i.e., to $\alpha_i Q_{-i} + 1$, in order to avoid shame. That is, she can change her ranking to $R_i = 1$. However, increasing Q_i decreases her earnings as, per definition, $\alpha_i Q_{-i} + 1 > Q_i^*$ for individuals with below-average ability ($A_i = 0$), i.e., she would work more than money maximizing, and $\pi_i(\alpha_i Q_{-i} + 1) < \pi_i(Q_i^*)$. Her utility can be described as

$$U_{i_{A_i=0,SA_i=1}}(\alpha_i Q_{-i} + 1) = \pi_i(\alpha_i Q_{-i} + 1) + s_{iTMT}. \tag{5}$$

An individual with low ability who stated high ability will increase output to $\alpha_i Q_{-i} + 1$ if $U_{i_{A_i=0,SA_i=1}}(\alpha_i Q_{-i} + 1) > U_{i_{A_i=0,SA_i=1}}(Q_i^*)$, i.e., if $\lambda_{iTMT} > \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$.[18] Put differently, the individual will increase output if the loss in image utility due to the feeling of shame is greater than the loss in utility due to the loss in earnings (from producing more output than money-maximizing).

While individuals with high ability will always state their true beliefs and maximize their earnings, individuals with low ability face a trade-off between earnings and image utility. They have to compare the utility from stating a low ability, defined in Eq. (3) to the utility from stating a high ability which is, depending on λ_{iTMT}, defined either in Eq. (4) or in Eq. (5).[19] If $\lambda_{iTMT} < \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$, the individual will state a high ability if $s_{iTMT} > \lambda_{iTMT}$. That is, if the image utility is higher than the loss in utility due to shame. If $\lambda_{iTMT} > \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$ the individual will state a high ability if $s_{iTMT} > \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$. That is, if the gain in image utility from stating high rather than low ability is higher than the loss in utility due to the loss in earnings from increasing output.[20]

Treatment effects related to the mode of the self-assessment.

Based on the above, the share of individuals stating a high ability increases in s_{iTMT} and decreases in λ_{iTMT}. In line with Ewers and Zimmermann (2015), I assume that image concerns matter more in treatments with public self-assessments than in treatments with private self-assessments, i.e., $s_{iAudience} = s_{iFeedback} \geq s_{iPrivate}$ and the dis-utility from shame is higher in treatments with public feedback compared to treatments with private feedback, i.e., $\lambda_{iFeedback} \geq \lambda_{iAudience} = \lambda_{iPrivate}$. Considering the treatment variations in the experiment, fewer individuals should state a high ability in *Private* than in *Audience* as $s_{iPrivate} \leq s_{iAudience}$. Note that λ_{iTMT} is constant across these two treatments as the feedback remains private. Also, fewer subjects should state a high ability in *Feedback* than in *Audience* as s_{iTMT} is constant across these treatments but the dis-utility from shame is higher in treatments with public feedback, i.e., $\lambda_{iFeedback} \geq \lambda_{iAudience}$. All hypotheses focus on the total share of subjects for generality, however, effects are expected to be driven by subjects with low ability, as explained above.

Hypothesis 1. *Confidence decreases when the self-assessment is private instead of public. The share of subjects stating a high ability is lower in Private than in Audience.*

Hypothesis 2. *Confidence decreases when subjects receive public instead of private feedback. The share of subjects stating a high ability is lower in Feedback than in Audience.*

As a second step, let us discuss gender differences. I expect men to have, on average, a lower α_i than women over all treatments, as there is ample evidence in the literature that men have a more optimistic self-assessment even controlling for ability (see Bengtsson et al., 2005; Dohmen and Falk, 2011; Möbius et al., 2011; Niederle and Vesterlund, 2007; Thoma, 2016). This means that men are more likely than women to believe that they are better than average in all treatments, and consequently, more men will state a high ability.

Hypothesis 3. *Men are, on average, more confident than women. The share of men stating a high ability is higher than the share of women stating a high ability.*

Concerning the effect of feedback, I hypothesize that women are more prone to the feeling of shame when the accuracy of the self-assessment is observed by others, following the findings of Ludwig et al. (2017). That is, the difference between $\lambda_{iFeedback}$ and $\lambda_{iAudience}$ is greater for women than for men.

[16] For simplicity, I assume that there is no loss in utility due to stating a low self-assessment. In addition, I exclude the possibility that individuals derive a positive utility from understating their ability, i.e., showing modesty. Relaxing this assumption does not qualitatively affect the predictions as long as the gain in utility from signaling high ability is larger than the gain from being modest.

[17] In case $s_{iTMT} = 0$ the individual is indifferent between both options, i.e., Eq. (2) = Eq. (3).

[18] In case $\lambda_{iTMT} = \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$, the individual is indifferent between both options.

[19] As there is no shame from understating one's ability the image utility from stating $SA_i = 0$ is the same for individuals with high and low ability.

[20] In case of equality the individual is indifferent between the two options, for all scenarios discussed above.

Hypothesis 4. *Women react more strongly to public feedback than men. The share of women stating a high ability decreases more between Audience and Feedback than the share of men stating a high ability.*

I add to the literature by investigating whether there also exist gender differences in s_{iTMT}, i.e., if there is a gender difference in how the image utility from stating a high ability is moderated by the observability of the self-assessment. To that end, I will compare how the shares of men and women who overstate their ability change between treatments. As the direction of the gender differences with respect to observability is not clear ex-ante, I test whether there is a difference, without forming a directional hypothesis.

Hypothesis 5. *There is a gender difference in the effect of a private instead of public self-assessment. The decrease in the share of subjects stating a high ability in Private compared to Audience varies by gender.*

Treatment effects related to the timing of the self-assessment. Concerning the timing of the self-assessment, subjects in *Ex post* cannot avoid shame by working harder. While this does not affect subjects with a high ability, subjects with a low ability will now only state a high ability if $s_{iTMT} > \lambda_{iTMT}$. In ex-ante treatments, however, subjects with low ability can work harder to live up to their self-assessment and will therefore also state a high ability if $\lambda_{iTMT} > \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$ and $s_{iTMT} > \pi_i(Q_i^*) - \pi_i(\alpha_i Q_{-i} + 1)$.[21] As self-assessments depend on s_{iTMT} and λ_{iTMT}, we can only compare treatments with constant image concerns and dis-utility from shame. Consequently, I compare the *Ex post* with the *Audience* treatment. In addition, I test for gender differences in the effect of the timing of the self-assessment. Lastly, I test if subjects actually work less without the ex-ante self-assessment.

Hypothesis 6. *Confidence decreases when the self-assessment is elicited after instead of before the task. The share of subjects stating a high ability decreases in Ex post compared to Audience.*

Hypothesis 7. *There is a gender difference in the effect of the timing of the self-assessment. The decrease in the share of subjects stating a high ability in Ex post compared to Audience varies by gender.*

Hypothesis 8. *Subjects work less when the self-assessment is elicited after instead of before the task. The deviation from the money-maximizing output is, on average, lower in Ex post than in Audience.*

4. Results

In this section, I first discuss how I empirically define overplacement.[22] Second, I compare self-assessments and the degree of overplacement across ex-ante treatments and gender (Hypotheses 1–5). Third, I compare self-assessments in the *Audience* and the *Ex post* treatment (Hypotheses 6 & 7). Lastly, I investigate whether subjects in the *Audience* treatment deviate more from money-maximizing output than subjects in *Ex post*, i.e., if there is a motivational effect of ex-ante self-assessments (Hypothesis 8). Unless indicated differently, reported p-values are based on bootstrapped score tests (999 replications) following logistic regressions with standard errors clustered at the session level.[23]

[21] In the experiment, subjects can improve their placement independent of the actions of the other participants as they are compared to an external reference group. See also Section 2.

[22] In the main part of the paper, I focus on overplacement as I also want to discuss a motivational effect of an overconfident ex-ante self-assessment. In Appendix D, I discuss how the treatments affect underconfidence. I find that all results for overplacing subjects are mirrored by underplacing subjects.

[23] I use bootstrapped score tests to derive the p-values to account for the small number of clusters, i.e., sessions (Cameron and Miller, 2015; Roodman et al., 2019). The results hold if I do not account for the small number of clusters, i.e., use clustered errors without bootstrapping, and if I use robust standard errors instead, the former are available from the author on request, the latter are presented in Appendix A.

Table 2
Self-assessments of men and women.

			Output higher than the average	
			No	Yes
Women:	Belief better than the average	No	62 (35%)	29 (16%)
		Yes	35 (20%)	50 (28%)
Men:	Belief better than the average	No	29 (18%)	18 (11%)
		Yes	37 (23%)	79 (48%)

Notes: Cells show the number of subjects. All treatments are included. Three subjects did not indicate their gender and are therefore excluded. Percentages are calculated within gender. The percentages for women do not add up to 100% because they are rounded to integer numbers.

To identify subjects who overplace, I compare subjects' relative self-assessments to their placement. In the main part of the paper, I use self-assessments as a binary variable with the value one if a subject stated that she is better than average, and compare it to placement in phase 1.[24] I use the comparison to the average instead of output quartiles, as this is the self-assessment that is made public and, therefore, most likely to be affected by the treatment variations. As a measure of placement, I use output, i.e., number of solved tasks, in phase 1, as it captures how much subjects work without the self-assessment. To account for session effects, I calculate the average output in phase 1 for each session separately.[25] If an individual completed more tasks than this average, she is classified as being better than average, and if she completed fewer tasks than average, she is classified as being worse than average, respectively. A person who stated in the self-assessment that she is better than average but her output is lower than average is classified as overplaced.

In total, about 59% of the subjects believe that they are better than the average subject. Comparing self-assessments to actual placement shows that about 36% of the subjects who believe that they are better than average are actually worse than average. Therefore, about 21% of the subjects overplace.

Fig. 1 shows the self-assessment (gray bars) as well as the overplacement (black bars) for each ex-ante treatment. To test hypotheses 1 & 2, I compare the share of subjects stating that they are better than average in the *Audience* treatment to the respective shares in the *Private* and the *Feedback* treatment (see also regressions in Table A.2 in Appendix A). The share of subjects who believe that they are better than average is not significantly lower when the self-assessment is private (62%) rather than public (69%) (*Private* vs. *Audience*, $p = 0.372$). However, the share is with 53% significantly lower when there is public feedback (*Feedback* vs. *Audience*, $p = 0.049$). The same pattern persists when controlling for ability by looking at subjects who overplaced. There is no significant decrease in overplacement in *Private* compared to *Audience* (21% overplace in *Private* compared to 31% in *Audience*, $p = 0.246$) but overplacement is with 16% significantly lower in *Feedback* than in *Audience* ($p = 0.075$). Therefore, I accept Hypothesis 2 but not Hypothesis 1. Although the effect of the *Private* treatment is not significant, these results are qualitatively in line with Ewers and Zimmermann (2015).

Before looking at gender specific treatment effects, let us discuss gender differences in confidence (Hypothesis 3). Table 2 shows self-assessments and actual performance by gender. Around 71% of men and 48% of women believe that they are better than the average ($p < 0.001$, test of proportions). However, due to gender differences in output (men complete an average of 91 tasks in phase 1 and women complete an average of 82 tasks, $p < 0.001$, two-sided t-test), there is no significant gender difference in overplacement as 20% of women and 23% of men overplace ($p = 0.526$, test of proportions).[26] Looking at each treat-

[24] I discuss other possible measures and the robustness of the results with respect to these alternative measures in Appendix B.

[25] Please note that in the experiment, the placement is based on the comparison to data from a previous experiment as was explained to the subjects.

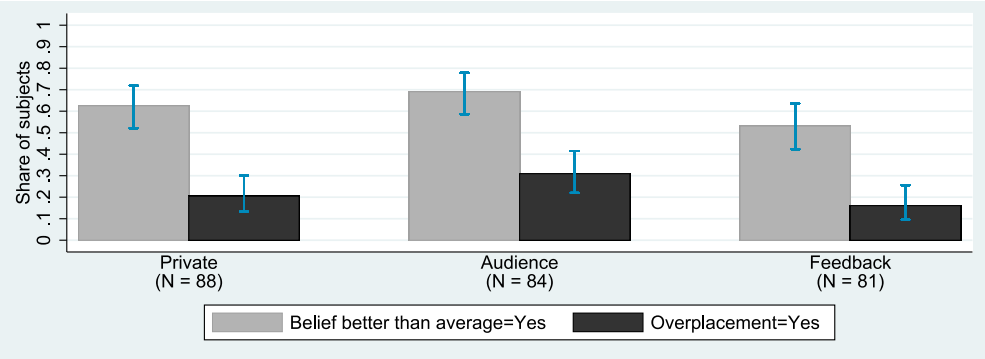


Fig. 1. Public feedback makes participants less confident. *Notes:* The figure shows the share of subjects that believe that they are better than average (gray bars) and the share of subjects who overplace (black bars). Whiskers show Wilson confidence intervals (Brown et al., 2001). Overplacement is one if a subject stated that she is better than the average but her performance in phase 1 was worse than the average. The figure includes the 253 subjects in the ex-ante treatments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

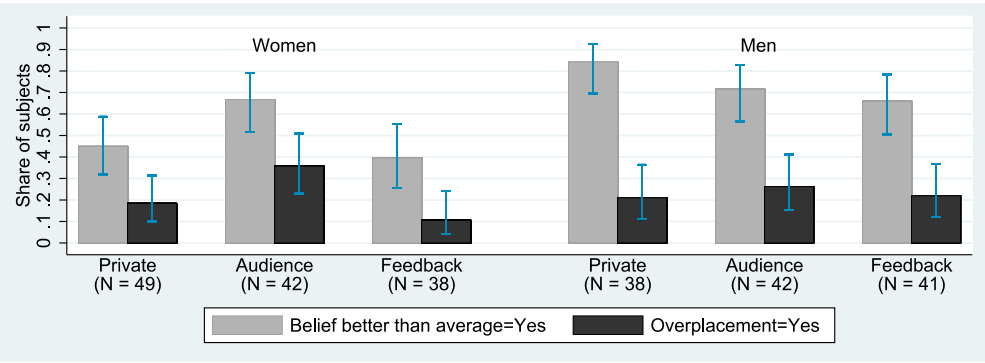


Fig. 2. Treatment effects are driven by women. *Notes:* The figure shows the share of subjects that believe that they are better than average (gray bars) and the share of subjects who overplace (black bars) by treatment and gender. Whiskers show Wilson confidence intervals (Brown et al., 2001). Overplacement is one if a subject stated that she is better than the average but her performance in phase 1 was worse than the average. The figure includes the 253 subjects in the ex-ante treatments, except for three subjects did not indicate their gender. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ment individually, shows that men are more confident than women in all individual treatments except for the *Audience* treatment, suggesting gender differences in treatment effects.[27] As for the pooled sample the differences become insignificant once I control for ability.[28]

Fig. 2 displays treatment effects by gender. The left panel shows the share of women who believe that they are better than the average subject (gray bars) and the share of overplacing women (black bars). The right panel shows both shares for men. The results reveal that only women drive the treatment effects observed for the pooled sample (see also regressions in Table A.2 in Appendix A). The share of women who believe that they are better than average increases in the social-signalling value of the self-assessment and is with 67% significantly higher in the *Audience* treatment than the 45% in the *Private* treatment ($p = 0.038$). Women are less confident when they face public feedback and the share of women stating high ability decreases to 40% in *Feedback* (*Feedback* vs. *Audience*: 40% vs. 67%, $p = 0.030$). The same pattern persists when controlling for ability by considering the share of women who overplace but the effect of the *Private* treatment is no longer significant (*Private* vs. *Audience*: 18% vs. 36%, $p = 0.186$ and *Feedback* vs. *Audience*: 11% vs. 36%, $p = 0.033$). For men, the treatments do not have the expected effects. There is no significant effect of the *Feedback* treatment and a marginally significant positive effect of the *Private* treatment. This positive effect, however, is in contrast to the hypothesis and not robust to different model specifications. I do not observe significant treatment effects for men when considering overplacement as the dependent variable.

Another way to control for ability is to include controls for performance. I use logistic regressions with standard errors clustered at the sessions level also controlling for cognitive and task-specific ability as well as risk aversion, which are reported in Table A.3 in Appendix A.[29] While the significant decrease in self-assessment for women when public feedback is introduced persists (*Feedback* vs. *Audience*, $p = 0.006$), I do not find a significant difference between women in the *Private* compared to the *Audience* treatment ($p = 0.331$). While the strong reaction of women to public feedback is in line with Ludwig et al. (2017), this paper additionally suggests that there are gender difference in the reaction to public self-assessment. However, the results on the public self-assessment are less robust than those on feedback. To get a more complete picture I will discuss how these findings relate to previous literature in the discussion.

To investigate whether the treatment effects differ between men and women, I use interaction effects in the logistic regressions in Table A.3. I start by considering the belief that someone is better than average as the dependent variable. When the self-assessment is private rather than public the odds that a woman will state that she is better than average decrease 6.9 times more than the odds that a man states that he is better than average ($p = 0.012$). As a result, Hypothesis 5 which states a gender difference in the effect of observability is accepted. Also, the negative effect of public feedback is stronger for women than for men. The odds of

[26] Approximately 41% of the 48% of women who believe that they are better than average were wrong, and 32% of the 71% of men who believe that they are better than average are wrong. Interestingly, even though men generate on average a higher output, 16% of women and only 11% of men state that they are worse than average when their performance is actually better than average, i.e., underplace ($p = 0.154$, test of proportions, see Appendix D for a detailed discussion of underplacement).

[27] Belief better than average: *Audience*: 67% of women vs. 71% of men, $p = 0.637$; *Private*: 45% of women vs. 84% of men, $p < 0.001$; *Feedback*: 40% of women vs. 66% of men, $p = 0.019$; *Ex post*: 43% of women vs. 64% of men, $p = 0.040$, test of proportions.

[28] Overplacement: *Audience*: 36% of women vs. 26% of men, $p = 0.245$; *Private*: 18% of women vs. 21% of men, $p = 0.754$; *Feedback*: 11% of women vs. 22% of men, $p = 0.171$; *Ex post*: 15% of women vs. 21% of men, $p = 0.423$, test of proportions.

[29] As before, I use bootstrapped score tests (999 replications) to derive the *p*-values reported in the text, to account for the small number of clusters, i.e., sessions. The results hold if I do not account for the small number of clusters, i.e., use clustered errors without bootstrapping and if I use robust standard errors instead, the former are available from the author on request, the latter are presented in Table 3. In addition, Table 4 shows, for better interpretability, OLS regressions with demeaned control variables.

stating a high ability decrease 3.9 times more for women than for men when public feedback is introduced ($p = 0.007$). Therefore, Hypothesis 4, i.e., women's self-assessments decrease more with public feedback than men's self-assessments, is accepted, suggesting that women are more averse to feelings of shame. Considering overplacement as the dependent variable does not show significant gender differences, and there is no significant explanatory power of the regressions. However, the effects go in the same direction as for the beliefs.

A second extension to previous literature is the timing of the self-assessment. To investigate whether the timing matters, I compare the self-assessments in ex-post and ex-ante treatments (see also regressions in Table A.5 and Fig. A.1 in Appendix A). I only consider treatments with public self-assessment and private feedback, i.e., *Audience* and *Ex post*. In these treatments the level of public exposure is constant. I find that eliciting beliefs ex post decreases the share of subjects who believe that they are better than average from 69% in *Audience* to 53% in *Ex post* (-16 percentage points, $p = 0.042$) indicating that the timing of the self-assessment matters and supporting Hypothesis 6. I find a similar pattern for overplacement, however, the treatment difference is not significant (18% in *Ex post* vs. 31% in *Audience*, i.e., -13 percentage points, $p = 0.127$). Interestingly, the treatment difference is again driven by women. The share of women who believe that they are better than average is 50% higher in *Audience* than in *Ex post* (*Ex post* vs. *Audience*: 43% vs. 67%, $p = 0.034$). Again a similar picture emerges for overplacement. The share of women who overplace in *Ex post* is less than half the size of the share of women who overplace in *Ex post* (*Ex post* vs. *Audience*: 15% vs. 36%, $p = 0.070$). To investigate whether the timing of the self-assessment affects men and women differently I use interaction terms (see Tables A.6 and A.7 in Appendix A). I find that the treatment effect on self-assessments is significantly larger for women than for men. The odds that a women states a high ability decrease 1.3 times more than the odds for a man when the self-assessment is elicited ex post rather than ex ante ($p = 0.006$). The same holds for overplacement, however there is overall no significant explanatory power of the regressions. Overall, I reject the null hypothesis that timing affects men's and women's self-assessment equally.

Robustness. In addition to the binary beliefs, I also elicit the subjects' belief distributions over performance quartiles. Subjects allocate probability points to each of the four performance quartiles, with quartile 1 indicating the highest performance quartile. These weighted average performance quartiles (WAQ) are highly correlated with the binary measure of beliefs (biserial correlation coefficient: 0.66, $p < 0.001$). In Table B.3 in Appendix B, I replicate the main results using the WAQ. The results are qualitatively similar to the results discussed above but less robust in terms of statistical significance. It is not surprising that the effects are weaker for the WAQ as the treatment variation targets the binary belief. In Appendix B, I additionally discuss the robustness of the main findings with respect to the definition of overplacement. The same pattern persists when I use money-maximizing output or average output across all sessions as a benchmark for ability, but the coefficients and regressions do not have significant explanatory power.

As the share of female subjects varies slightly between sessions and previous literature investigating gender differences in competitiveness has shown that the willingness to enter competition is responsive to the gender-composition of the sample (see for example Gneezy et al., 2003), the gender-composition of the sessions might also affect men's and women's confidence. I repeat the analysis presented above, controlling for the gender-composition in Appendix A (see Table A.8). I find that the treatment differences discussed above do not change if I control for the share of women in the sessions.

Does the degree of overplacement matter? As overplacement is defined as a binary variable, I have, so far, not accounted for the degree to which a subject is overstimating his or her performance. To gain deeper insights on who is affected most by the treatments, I use different output cut-offs to define subjects as overplacing, in Appendix C. The results provide suggestive evidence that the treatment effects are mainly driven by subjects whose performance is close to average performance.

Table 3
Timing does not affect deviation from money-maximizing output.

	Average deviation from money-maximizing output		
	(1) Pooled	(2) Women	(3) Men
Ex post	–0.130	3.547	–4.014
	(3.505)	(4.637)	(5.284)
	[0.961]	[0.626]	[0.313]
Male	–2.839		
	(3.528)		
	[0.472]		
Constant	7.421**	5.479*	6.524*
	(2.924)	(3.292)	(3.347)
R^2	0.00	0.01	0.01
N	173	89	84

Notes: OLS regressions with the deviation from money-maximizing output in phase 2 as the dependent variable and *Audience* as the baseline. Money-maximizing output is estimated for each individual based on the speed at which subjects solve the tasks in phase 2 and estimate the output at which they should switch to being idle to maximize their earnings (see Haeckl et al., 2018, for a detailed discussion of this proxy for money-maximizing output). I only consider treatments with public self-assessment and private feedback, i.e., *Audience* and *Ex post*, which leads to 173 observations. In columns (2)-(3), I split the sample by gender. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on wild cluster bootstrap *t*-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

To sum up, while I find that women adapt their self-assessment depending on the social-signalling value and the timing of the self-assessments, men do not seem to adapt their self-assessment across treatments. I interpret the results as evidence that women are more receptive to social image concerns than men. Precisely, women want to signal high ability and (consciously) overstate their ability as long as their self-assessments are not verifiable, i.e., in the *Audience* and are more receptive to the feeling of shame when self-assessments are publicly verifiable, i.e., in the *Feedback* treatment (see also Ludwig et al., 2017). In addition, women are more confident when they can influence the accuracy of the self-assessment ex post, i.e., in the *Audience* compared to the *Ex Post* treatment.

If inflated self-assessments are conscious, they could motivate subjects to work harder to meet their self-assessment (Hypothesis 8). As subjects spend, on average, 539 out of 1200 seconds working on the task in phase 1 (median 531 seconds), there is scope for an increase in effort. To test, I compare average deviations from the money-maximizing output in phase 2 between *Audience* and *Ex post* in Table 3.

I do not find evidence that eliciting self-assessments ex post makes subjects work less as the timing of the self-assessment does not explain the variation in deviation from the money-maximizing benchmark (see variable *Ex post*).[30] Based on the model, subjects who overstate their ability ex ante should work harder to live up to their self-assessment leading to higher output in *Audience* than in *Ex post*. While the share of subjects overstating their ability does not significantly differ between treatments for men, women stated higher self-assessments in ex-ante treatments. Consequently, we would expect larger effects for women. A subgroup analysis only considering women (see column(2)) does not support this hypothesis. The treatment effect is not significant and the estimate does not have the expected sign.

[30] In general, subjects are, on average, money maximizers and do not significantly deviate from the money-maximizing number of solved tasks. A one-sample *t*-test testing if the deviation from money maximizing behavior in phase 1 does not reject the hypothesis that the average deviation is 0 ($p = 0.240$).

Another way to look at behavioral effects of overplacement is to check for differences in the average change in work time between phases 1 and 2. As subjects have the option to stop working and remain idle, work time is also a proxy for effort. I do not find significant treatment differences in the change in work time between *Audience* and *Ex post* (see Table A.9 in Appendix A). Lastly, the null effect could also mask heterogeneous effects based on ability. In Table A.10, I split subjects into two groups based on their performance in phase 1 and test if the treatment effects vary between groups. I do not find evidence that heterogeneous treatment effects are causing the null result. I, therefore, reject Hypothesis 8.

5. Conclusion and discussion

I use a real-effort experiment to investigate how variations in the image concerns related to self-assessments affect men's and women's stated confidence. To cleanly identify the effect of image concerns, I use a setting without any strategic incentives. I find that women increase their self-assessment when it is made public and decrease it when they receive public feedback. I do not find such effects for men. These results are consistent with an explanation according to which women face public pressure to appear in a rosy light and adapt their self-assessments in response to social-image concerns.

To put the results of this paper into perspective, I compare them to papers sharing the core design feature of treatment manipulation of social image concerns of self-assessments on an objective performance scale. Three papers are suitable for this comparison, namely Ewers and Zimmermann (2015), short EZ, Schwardmann and van der Weele (2019), short SV, and Ludwig et al. (2017), short LFT (see Appendix E). Considering the findings together shows that people's confidence is higher if they have to conduct their self-assessments publicly rather than privately and lower when the accuracy of the self-assessment is observable by a third person. This pattern exists for women in all papers. However, the increase in confidence between private and public self-assessment varies between +4 percentage points (pp) in SV and +18pp in this paper. The decrease in women's confidence is more robust ranging between -19pp in LFT and -25pp in this paper. All papers also find an increase in confidence from private to public self-assessment for men, however, it ranges between +4pp in SV to +20pp in EZ. The results on how men's confidence is affected by public feedback are less conclusive and range from +3pp in LFT to -13pp in EZ.

In general, the results in this paper are, as expected, very similar to EZ, with one exception. While the increase in confidence in the *Audience* treatment is driven by women in this experiment, it is driven by men in EZ. A potential explanation for this difference is the different timing of the self-assessments in both experiments. While self-assessments take place before the real-effort task in this experiment they take place ex post in EZ. The possibility to affect performance after the self-assessment might affect women's confidence. This is also in line with comparisons between the *Audience* and the *Ex post* treatment in this paper. The share of women who overplace decreases by 19pp if beliefs are elicited after rather than before the real-effort task.

The suggestion that gender differences in confidence are sensitive to characteristics of the self-assessment is also in line with previous literature finding that gender differences in confidence vary with the type of task used or the comparison group (Bordalo et al., 2019; Coffman, 2014; Dreber et al., 2014; Ring et al., 2016). These differences could even be amplified in a setting with social image concerns as these have also been found to be heavily depending on the context and peer group (Bursztyn et al., 2018). While in some groups women might be expected to be modest, in others they might be expected to signal capability. It is plausible that among university students at a business school, women are expected to be good at math and follow the social pressure to signal their high ability. Another potential explanation is that women who select into studying at a business university are less modest (see for example Hardies et al., 2013; Nekby et al., 2008, for a discussion of selection and overconfidence).

I also test for a motivational effect of overconfidence. The idea that people increase effort to live up to a prior statement is based on two well-developed findings in the literature. First, it is plausible that people derive an intrinsic value of a positive self-image through the consumption value of overconfidence, as discussed by Bénabou and Tirole (2002) or through ego-utility as in Köszegi (2006). Second, findings in goal theory (Goerg and Kube, 2012; Koch and Nafziger, 2011; 2020; Latham and Locke, 1991; Smithers, 2015) suggest that people are motivated to increase effort if they are provided with goals, even if goals are non-incentivized. In this experiment, subjects' self-assessments can be interpreted as reference points, and subjects can increase their output to avoid negative feedback. However, I do not find evidence for such a motivational effect.

The null result seems to be in contrast with Chen and Schildberg-Hörisch (2019), who find that effort provision increases in overconfidence, but it is not. In their setting, subjects do not know when they have reached their money-maximizing output and the motivational value of overconfidence stems from subjects who overestimate their return to effort. In contrast, in this experiment, subjects know their return to effort as they receive continuous feedback on the time it takes them to complete each screen. Therefore, the motivational effect stems from subjects who overstate their placement for image concerns and are willing to work more to live up to their statement.

While I find that women are more confident when they have the opportunity to affect their placement ex post, i.e., they are more confident in *Audience* than in *Ex post*, they do not increase effort in order to live up to their self-assessment making it puzzling that they increased their self-assessment in the first place. Below, I first discuss two potential design-related explanations and second discuss other mechanisms that might be at play.

The first potential design-related explanation is that it is not possible to increase effort. This is unlikely as the real-effort task was designed to allow subjects to improve performance by working longer and the average work time in phase one is 539 out of 1200 seconds, indicating that there is room to increase effort.[31]

The second potential design-related explanation is that I do not have sufficient power to detect changes in effort, especially as only subjects with a below-average performance can be overconfident and thereby motivated by the self-assessment. Only considering the 173 participants in treatments with public self-assessment, I can detect an effect of 0.43 standard deviations with a power of 80% and $\alpha = 0.05$. This is comparable to effect sizes for example found in Smithers (2015), who set a non-incentivized goal at median performance (hence, only allowing subjects who perform worse than the median to be motivated).

Concerning other mechanism that might mute the effect, a potential explanation is that the shift in self-assessment is not conscious. One necessary assumption for subjects to increase effort after an overconfident public self-assessment is that they are aware (at least to some extent) that they have been overstating their ability. However, as for example discussed in von Hippel and Trivers (2011), subjects might subconsciously deceive themselves. This is also in line with Schwardmann and van der Weele (2019) finding that subjects' confidence in a private and incentivized self-assessment increases if they ex post have to convince others of their high ability.[32] This mechanism, however, does not explain why subjects in *Ex post* are less confident than in *Audience* as the social signalling value of the self-assessment is held constant between

[31] In addition, the task has been used previously to detect changes in effort between treatments (Haeckl et al., 2018).

[32] The authors provide evidence that this behavior is neither driven by lying aversion nor preferences for consistency, but is strongest for subjects believing that it is easier to persuade others when one is confident.

these treatments. A potential explanation is that there is less ambiguity about their relative placement after the real-effort task. While subjects have experience in working on the real-effort task in both treatments (from phase 1), there might be ambiguity about the performance in the upcoming phase leaving more scope for self-deception in *Audience* than in *Ex post*.[33]

Lastly, present bias may help to understand the results. It is possible that subjects underestimate the effort cost from working and by that overestimate their willingness to work hard to live up to their self-assessment. If a subject is naive about her present bias and cares about her social image, she will increase her public self-assessment when it is elicited prior to the real-effort task under the expectation that she will work harder in the real-effort task. Once she works on the real-effort task she understands the true cost of effort and does not live up to her self-assessment.[34] While this paper cannot identify the different potential mechanisms, it opens an interesting avenue for future research.

Declaration of Competing Interest

None.

Appendix A. Additional Tables

Table A.1
Treatments.

Treatment	Observations	Share women
Ex post	89	53.33%
Private	88	56.82%
Audience	84	50.00%
Feedback	81	47.50%

Table A.2
Self-assessment and overplacement vary across treatments and gender.

	Logistic regression (Odds Ratios)					
	Belief better than average = Yes			Overplacement = Yes		
	(1) Pooled	(2) Women	(3) Men	(4) Pooled	(5) Women	(6) Men
Private	0.747	0.407**	2.133	0.574	0.405*	0.752
	(0.242)	(0.178)	(1.202)	(0.204)	(0.199)	(0.400)
	[0.372]	[0.038]	[0.095]	[0.246]	[0.186]	[0.461]
Feedback	0.507**	0.326**	0.771	0.426**	0.212**	0.793
	(0.165)	(0.153)	(0.368)	(0.164)	(0.132)	(0.410)
	[0.049]	[0.030]	[0.336]	[0.075]	[0.033]	[0.648]
Constant	2.231***	2.000**	2.500***	0.448***	0.556*	0.355***
	(0.528)	(0.657)	(0.857)	(0.106)	(0.180)	(0.125)
Pseud. R^2	0.10	0.04	0.03	0.02	0.06	0.00
	OLS regression					
	Belief better than average = Yes			Overplacement = Yes		
	(1) Pooled	(2) Women	(3) Men	(4) Pooled	(5) Women	(6) Men
Private	–0.065	–0.218**	0.128	–0.105	–0.173*	–0.051
	(0.073)	(0.103)	(0.093)	(0.067)	(0.093)	(0.096)
	[0.438]	[0.038]	[0.105]	[0.276]	[0.189]	[0.529]
Feedback	–0.160**	–0.272**	–0.056	–0.149**	–0.252***	–0.042
	(0.075)	(0.109)	(0.103)	(0.065)	(0.090)	(0.095)
	[0.055]	[0.029]	[0.409]	[0.093]	[0.040]	[0.700]
Constant	0.690***	0.667***	0.714***	0.310***	0.357***	0.262***
	(0.051)	(0.074)	(0.071)	(0.051)	(0.075)	(0.069)
R^2	0.02	0.05	0.03	0.02	0.06	0.00
N	253	129	121	253	129	121

Notes: The upper panel shows odds ratios from logistic regressions and the lower panel shows an OLS regression.The baseline is the *Audience* treatment. All 253 subjects in the ex-ante treatments are considered in columns (1) and (4) while the sample is split by gender in the other columns. Three subjects did not indicate their gender and are therefore excluded in the respective columns. Numbers in parentheses show robust standard errors. Numbers in brackets show *p*-values based on cluster bootstrapped score tests (999 replications) in the upper sample and wild cluster bootstrap *t*-tests (999 replications) in the lower panel to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

[33] For example Robinson and Ryff (1999) discuss that people are more prone to self-deception when thinking about future rather than past happiness and Gneezy et al. (2020) show how self-deception varies with ambiguity in a setting where subjects engage in selfish behavior.

[34] This line of reasoning is also in line with recent literature in political economics showing that present biased individuals are less likely to vote even after having stated that they intend to do so (Hill, 2020).

Table A.3
There are gender-differences in treatment effects.

	Belief better than average = Yes		Overplacement = Yes	
	(1)	(2)	(3)	(4)
Private	0.407**	0.580	0.405*	0.401*
	(0.178)	(0.279)	(0.199)	(0.199)
	[0.039]	[0.331]	[0.186]	[0.186]
Feedback	0.326**	0.242***	0.212**	0.211**
	(0.152)	(0.132)	(0.131)	(0.133)
	[0.030]	[0.006]	[0.030]	[0.034]
Male	1.250	0.607	0.639	0.644
	(0.593)	(0.325)	(0.305)	(0.311)
	[0.4390]	[0.056]	[0.127]	[0.241]
Private_x_Male	5.236**	6.885**	1.856	1.901
	(3.726)	(5.244)	(1.342)	(1.379)
	[0.006]	[0.012]	[0.306]	[0.300]
Feedback_x_Male	2.366	3.933*	3.743	3.799
	(1.577)	(2.886)	(3.021)	(3.100)
	[0.015]	[0.007]	[0.139]	[0.142]
Risk Aversion		0.994		0.969
		(0.060)		(0.072)
		[0.954]		[0.696]
Ability		1.106***		
		(0.027)		
		[0.001]		
Cognitive Ability		1.027		0.999
		(0.037)		(0.032)
		[0.472]		[0.966]
Constant	2.000**	0.005***	0.556*	0.648
	(0.656)	(0.007)	(0.179)	(0.510)
Pseudo R^2	0.08	0.16	0.03	0.03
N	250	250	250	250

Notes: Coefficients are reported as odds ratios. The baseline group are females in the *Audience* treatment. All 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. Ability is proxied by output generated in the five-minute piece-rate phase. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on cluster bootstrapped score tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

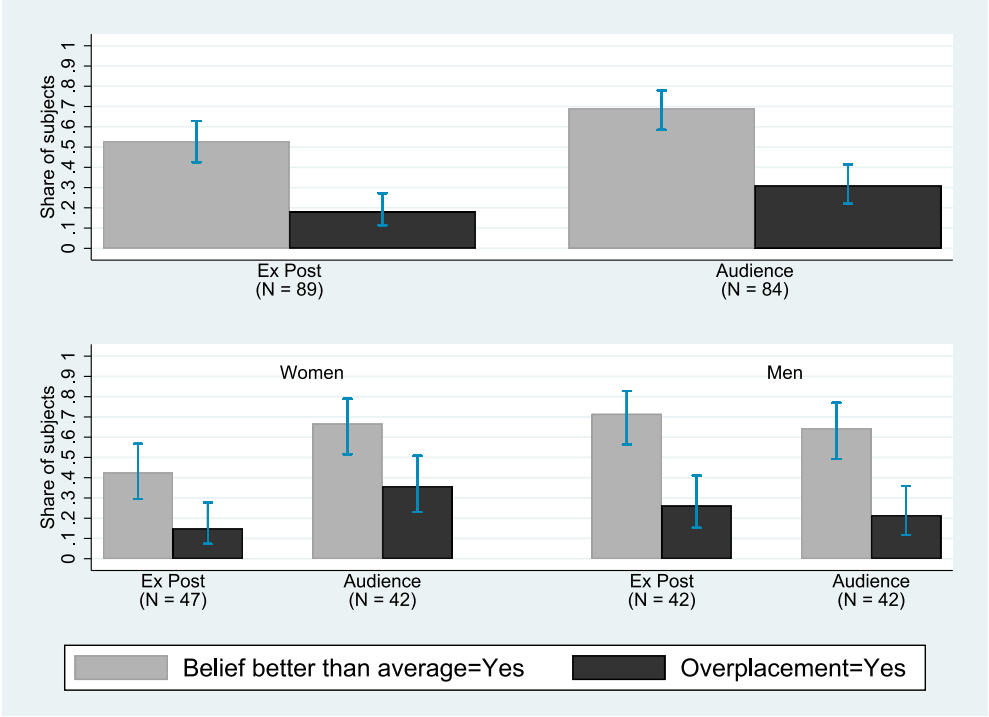


Fig. A.1. Treatment effects of the timing of the self-assessment are driven by women. *Notes:* The figure shows the share of subjects that believe that they are better than average (gray bars) and the share of subjects who overplace (black bars) depending on the timing of the self-assessment (upper panel) and split by gender (lower panel). Whiskers show Wilson confidence intervals (Brown et al., 2001). Overplacement is one if a subject stated that she is better than the average but her performance in phase 1 was worse than the average. The figure includes the 173 subjects in the treatments with public self-assessments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table A.4
Self-assessment and overplacement vary across treatments (OLS).

	Belief better than average = Yes (1)	Overplacement = Yes (2)
Private	–0.132	–0.175*
	(0.101)	(0.094)
	[0.289]	[0.191]
Feedback	–0.301***	–0.253***
	(0.109)	(0.092)
	[0.016]	[0.043]
Male	–0.096	–0.094
	(0.103)	(0.103)
	[0.150]	[0.199]
Private_x_Male	0.368***	0.126
	(0.134)	(0.135)
	[0.017]	[0.281]
Feedback_x_Male	0.285*	0.212
	(0.145)	(0.132)
	[0.014]	[0.156]
Risk Aversion	–0.001	–0.005
	(0.012)	(0.012)
	[0.939]	[0.689]
Ability	0.019***	
	(0.004)	
	[0.002]	
Cognitive Ability	0.006	–0.000
	(0.007)	(0.006)
	[0.456]	[0.968]
Constant	0.683***	0.357***
	(0.074)	(0.075)
R^2	0.20	0.03
N	250	250

Notes: OLS regressions with demeaned control variables for easier interpretability replicating the results of Table A.3. The baseline group are females in the *Audience* treatment. All 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. Ability is proxied by output generated in the five-minute piece-rate phase. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on wild cluster bootstrap *t*-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.5
Self-assessment and overplacement vary with the timing of the assessment and gender.

	Logistic regression (Odds Ratios)					
	Belief better than average = Yes			Overplacement = Yes		
	(1) Pooled	(2) Women	(3) Men	(4) Pooled	(5) Women	(6) Men
Ex post	0.502**	0.370**	0.720	0.489**	0.315**	0.769
	(0.160)	(0.164)	(0.340)	(0.178)	(0.165)	(0.398)
	[0.042]	[0.034]	[0.239]	[0.127]	[0.070]	[0.534]
Constant	2.231***	2.000**	2.500***	0.448***	0.556*	0.355***
	(0.528)	(0.658)	(0.859)	(0.106)	(0.180)	(0.125)
Pseud. R^2	0.02	0.04	0.00	0.02	0.05	0.00
	OLS regression					
	Belief better than average = Yes			Overplacement = Yes		
	(1) Pooled	(2) Women	(3) Men	(4) Pooled	(5) Women	(6) Men
Ex post	–0.162**	–0.241**	–0.071	–0.130**	–0.208**	–0.048
	(0.074)	(0.104)	(0.103)	(0.065)	(0.091)	(0.094)
	[0.049]	[0.034]	[0.304]	[0.160]	[0.075]	[0.592]
Constant	0.690***	0.667***	0.714***	0.310***	0.357***	0.262***
	(0.051)	(0.074)	(0.071)	(0.051)	(0.075)	(0.069)
R^2	0.03	0.06	0.01	0.02	0.06	0.00
N	173	89	84	173	89	84

Notes: The upper panel shows odds ratios from logistic regressions and the lower panel shows an OLS regression.The baseline is the *Audience* treatment. All 173 subjects in treatments with public self-assessments and private feedback, except one subject who faced technical problems, are considered in columns (1) and (4) while the sample is split by gender in the other columns. Numbers in parentheses show robust standard errors. Numbers in brackets show *p*-values based on cluster bootstrapped score tests (999 replications) in the upper sample and wild cluster bootstrap *t*-tests (999 replications) in the lower panel to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.6
The effect of the timing of the assessment varies by gender.

	Belief better than average = Yes		Overplacement = Yes	
	(1)	(2)	(3)	(4)
Ex post	0.370**	0.419*	0.315**	0.298**
	(0.164)	(0.207)	(0.165)	(0.161)
	[0.033]	[0.111]	[0.070]	[0.072]
Male	1.250	0.594	0.639	0.584
	(0.593)	(0.317)	(0.305)	(0.284)
	[0.328]	[0.021]	[0.127]	[0.107]
Ex post x Male	1.944	3.791*	2.440	2.734
	(1.255)	(2.748)	(1.792)	(2.052)
	[0.010]	[0.007]	[0.020]	[0.034]
Risk Aversion		0.946		0.965
		(0.075)		(0.080)
		[0.384]		[0.734]
Ability		1.115***		
		(0.030)		
		[0.007]		
Cognitive Ability		1.018		1.035
		(0.038)		(0.035)
		[0.692]		[0.310]
Constant	2.000**	0.005***	0.556*	0.326
	(0.657)	(0.007)	(0.179)	(0.264)
Pseudo R^2	0.04	0.15	0.03	0.03
N	173	173	173	173

Notes: Coefficients are reported as odds ratios. The baseline group are females in the *Audience* treatment. All 173 subjects in treatments with public self-assessments and private feedback are included. Ability is proxied by output generated in the five-minute piece-rate phase. Numbers in parentheses indicate robust standard errors. Numbers in brackets show p-values based on cluster bootstrapped score tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.7
Self-assessment and overplacement vary with the timing of the assessment (OLS).

	(1) Belief better than average = Yes	(2) Overplacement = Yes
Ex post	0.829*	0.805**
	(0.085)	(0.075)
	[0.101]	[0.092]
Male	0.900	0.896
	(0.092)	(0.092)
	[0.042]	[0.184]
Ex post x Male	1.318**	1.196
	(0.183)	(0.158)
	[0.006]	[0.019]
Risk Aversion	0.988	0.993
	(0.016)	(0.015)
	[0.350]	[0.738]
Ability	1.021***	
	(0.004)	
	[0.005]	
Cognitive Ability	1.004	1.006
	(0.007)	(0.006)
	[0.648]	[0.308]
Constant	1.979***	1.439***
	(0.147)	(0.109)
R^2	0.19	0.04
N	173	173

Notes: OLS regressions with demeaned control variables for easier interpretability replicating the results of Table A.6 for simpler interpretability. The baseline group are females in the *Audience* treatment. All 173 subjects in treatments with public self-assessments and private feedback are included. Ability is proxied by output generated in the five-minute piece-rate phase. Numbers in parentheses indicate robust standard errors. Numbers in brackets show p-values based on wild cluster bootstrap t-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.8
Replication of Table A.3 controlling for gender composition.

	(1) Belief better than average = Yes	(2) Overplacement = Yes
Private	0.629	0.512
	(0.337)	(0.284)
	[0.541]	[0.302]
Feedback	0.232***	0.181***
	(0.130)	(0.117)
	[0.009]	[0.019]
Male	0.606	0.624
	(0.326)	(0.307)
	[0.060]	[0.187]
Private x Male	6.883**	1.948
	(5.265)	(1.418)
	[0.013]	[0.310]
Feedback x Male	3.890*	3.848
	(2.858)	(3.156)
	[0.007]	[0.142]
Risk Aversion	0.996	0.977
	(0.060)	(0.073)
	[0.974]	[0.776]
Ability	1.106***	
	(0.027)	
	[0.002]	
Cognitive Ability	1.025	0.993
	(0.037)	(0.032)
	[0.528]	[0.843]
Share women	0.987	0.960
	(0.039)	(0.037)
	[0.607]	[0.306]
Constant	0.010*	5.490
	(0.025)	(11.750)
Pseudo R^2	0.16	0.04
N	250	250

Notes: Coefficients are reported as odds ratios. The baseline group are females in the *Audience* treatment. All 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. Ability is proxied by output generated in the five-minute piece-rate phase. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on cluster bootstrapped score tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.9
There is no treatment variation in the change in work time.

	Average change in work time between phases		
	(1) Pooled	(2) Women	(3) Men
Ex post	10.705	58.608	–39.890
	(34.525)	(59.224)	(33.072)
	[0.861]	[0.524]	[0.314]
Male	-8.789		
	(34.043)		
	[0.824]		
Constant	45.928	20.630	62.436***
	(34.689)	(42.894)	(21.627)
R^2	0.00	0.01	0.02
N	173	89	84

Notes: OLS regressions with the change in work time between phases 1 and 2 as the dependent variable and *Audience* as the baseline. I only consider treatments with public self-assessment and private feedback, i.e., *Audience* and *Ex post*, which leads to 173 observations. In columns (2)-(3), I split the sample by gender. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on wild cluster bootstrap *t*-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.10
Behavioral change between ex-ante and ex-post treatments by performance in phase 1.

	(1) Change in Worktime	(2) Deviations from money-max. output
Ex post	–26.989	0.333
	(46.098)	(4.566)
	[0.646]	[0.952]
Above median ability	–162.499***	10.786**
	(44.385)	(4.629)
	[0.116]	[0.431]
Ex post x Above median ability	90.152	–1.112
	(66.038)	(6.905)
	[0.065]	[0.892]
Constant	111.175***	1.379
	(31.506)	(2.934)
R^2	0.08	0.05
N	173	173

Notes: OLS regressions with the change in work time between phases 1 and 2 as the dependent variable in column (1) and the deviation from money-maximizing output in phase to in column (2). The baseline group are subjects in the *Audience* treatment. I only consider treatments with public self-assessment and private feedback, i.e., *Audience* and *Ex post* which leads to 173 observations. The variable *Above median ability* is a dummy variable with value one if the subject's performance in phase 1 was above the median performance in her session. Numbers in parentheses indicate robust standard errors. Numbers in brackets show p-values based on wild cluster bootstrap t-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Appendix B. Robustness

Overplacement based on money-maximizing output in phase 2. In the main part of the paper, the measure of overplacement is based on output in phase 1. The reason is that output in phase 1 is utility-maximizing, also controlling for intrinsic motivation to solve the task and effort cost, i.e., it accounts for subjects' preferences for working on the task compared to being idle, assuming subjects are rational. Thus, the measure is less restrictive than other potential measures like the benchmark for money-maximizing output or ability measured in the piece rate phase. However, overplacement is usually defined as an over-estimation of one's relative ability. To that end, I use the benchmark for money-maximizing output, as defined in Section 2, in phase 2 to generate a second measure for overplacement.[35] Subjects are now classified as overplaced if they state that they are better than average, but their money-maximizing output is lower than average. Using this measure for overplacement changes the classification for 20% of the subjects. More precisely, 25 subjects who have been classified as overplacing based on their placement in phase 1, are now classified as well-calibrated, and 45 subjects who have been well-calibrated are now classified as overplaced. This asymmetric change can be explained by intrinsic motivation. Subjects, on average, deviate positively from the money-maximizing benchmark, letting them appear more capable. This decreases the likelihood that participants are classified as overplaced.

Table B.1 replicates Table A.3 using the money-maximizing output as the benchmark for overplacement. The p-values reported below are based on cluster bootstrapped score tests (999 replications) also shown in the table. The odds that women overplace are not significantly lower when the self-assessment is private (odds ratio: 0.534, $p = 0.178$) and decrease significantly when public feedback is introduced (odds-ratio: 0.319, $p = 0.028$). While, the gender difference in the reaction to private rather than public self-assessments is not significant ($p = 0.114$), there is a significant gender difference in the reaction to public feedback ($p = 0.022$). Over all, the size and direction of all coefficients is similar to Table A.3.

Overplacement based on overall performance in phase 1. In the main part of the paper, I use session averages as the performance cut-off to define overplacement. I do so to account for randomly occurring session-specific differences.

As a robustness check I replicate the main results using the average performance across all sessions as the cut-off to define overplacement instead in Table B.2. I find a significant decrease in women's confidence in the presence of public feedback and when the self-assessment is stated after rather than before the real-effort task. While there are no significant gender differences, the coefficients are qualitatively similar to the main results. As for columns (3) and (4) in Table A.3, the variables have jointly no explanatory power.

Self-assessment measured as weighted beliefs. While I focus on the "better-than-average" self-assessment in the main part of the paper, Table B.3 shows that the gender differences in treatment effects persist if I use weighted beliefs based on the output-quartile assessment.

The dependent variable is the weighted average-output quartile (WAQ), which results from the probability points a subject allocated to each of the four quartiles. A WAQ of 1 means that a subject stated that she is 100% sure that she is in the top 25% of the participants, a WAQ of 4 means that she is 100% sure that she is among the bottom 25%, respectively. The average WAQ is 2.21, and the standard deviation is 0.55. Fig. B.1 shows how the WAQ relates to the binary belief measure. The WAQ of subjects who state that they believe that they are better than the average is 0.78 points lower than the WAQ of subjects who state that they believe that they are worse than the average (1.92 vs. 2.70, $p < 0.001$, two-sided t-test). Another way to compare the two groups is to look at the number of probability points they put in each quartile. Subjects who state that they believe

[35] See also Haeckl et al. (2018) for a discussion of the proxy for money-maximizing output.

Table B.1
Overplacement based on money-maximizing output.

	Overplacement = Yes	
	(1)	(2)
Private	0.533	0.534
	(0.237)	(0.241)
	[0.167]	[0.178]
Feedback	0.301**	0.319**
	(0.157)	(0.169)
	[0.033]	[0.028]
Male	0.473	0.534
	(0.223)	(0.255)
	[0.073]	[0.094]
Private x Male	2.153	2.081
	(1.444)	(1.400)
	[0.105]	[0.114]
Feedback x Male	3.019	2.787
	(2.198)	(2.037)
	[0.031]	[0.022]
Risk Aversion		0.967
		(0.059)
		[0.635]
Cognitive Ability		0.955
		(0.030)
		[0.285]
Constant	0.750	2.113
	(0.234)	(1.568)
Pseudo R^2	0.02	0.03
N	250	250

Notes: Coefficients are reported as odds ratios. The baseline group are women in the *Audience* treatment. All 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. Overplacement is based on money-maximizing output in phase 2. Numbers in parentheses indicate robust standard errors. Numbers in brackets show p-values based on cluster bootstrapped score tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

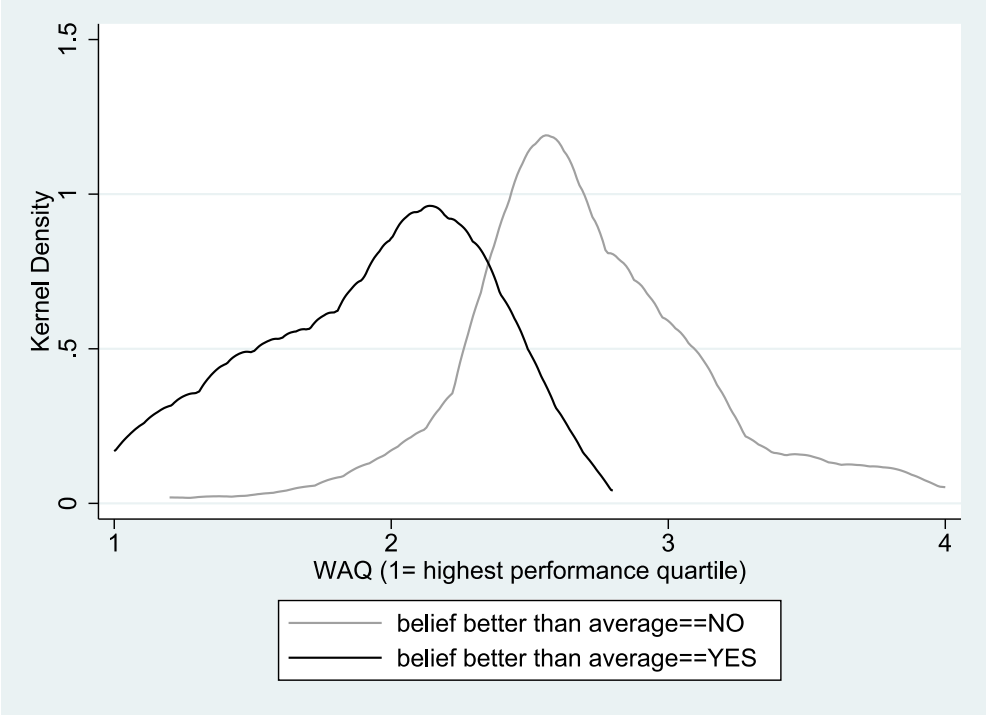


Fig. B.1. WAQ and the binary self-assessment are highly correlated. *Notes:* The figure shows the WAQ for subjects who belief that they are better than the average (black line) and subjects who belief that they are worse than the average (gray line). A WAQ of 1 indicates that a person believes that she is in the top 25% of participants with certainty (She put 100 probability points in the top quartile), a WAQ of 4 means that a person is certain that she is among the worst 25%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table B.2
Overplacement based on average performance in phase 1.

	Overplacement = Yes	
	(1)	(2)
Private	0.894	
	(0.423)	
	[0.802]	
Feedback	0.374	
	(0.224)	
	[0.098]	
Male	1.118	1.043
	(0.539)	(0.506)
	[0.686]	[0.841]
Private x Male	1.665	
	(1.109)	
	[0.417]	
Feedback x Male	1.701	
	(1.330)	
	[0.554]	
Ex post		0.335*
		(0.194)
		[0.013]
Ex post x Male		1.930
		(1.495)
		[0.211]
Risk Aversion	0.968	0.917
	(0.067)	(0.081)
	[0.626]	[0.443]
Cognitive Ability	1.003	1.033
	(0.031)	(0.035)
	[0.882]	[0.247]
Constant	0.429	0.294
	(0.329)	(0.245)
Pseudo R^2	0.03	0.04
N	250	173

Notes: Coefficients are reported as odds ratios. The baseline group are women in the *Audience* treatment. In column (1), all 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. In column (2), only subjects in treatments with public self-assessment, i.e., *Audience* and *Ex post*, are included which leads to 173 observations. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on cluster bootstrapped score tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

that they are better than the average put on average 78 probability points into the top 2 quartiles while subjects who believe that they are worse than the average put on average 40 probability points into the top 2 quartiles ($p < 0.001$, two-sided *t*-test).

Column (1) of Table B.3 shows that the difference in WAQ between private and public self-assessment is significantly smaller for men than for women (-0.488, $p = 0.014$, see *Private x Male* in column 1), i.e., their confidence decreases less due to the lack of observability in the *Private* treatment. In line with the findings in Table A.3, we see that women's self-assessment gets more pessimistic in the presence of public feedback, however, this change is no longer statistically significant once I account for the small number of clusters by using the wild bootstrap (0.275, $p = 0.159$). Similarly, the gender difference in the effect of feedback has the expected sign but is not statistically significant ($p = 0.392$).

Column (2) replicates the results presented in Table A.6 and investigates whether the timing of the self-assessment affects confidence. While the coefficient has the expected sign, i.e., women are less confident when the self-assessment happens after the task rather than before, the treatment effect is not statistically significant ($p = 0.490$). The same holds for the gender difference in the treatment effect.

Table B.3
Self-assessment varies across treatments (WAQ).

	Weighted average-output quartile	
	(1)	(2)
Private Private	0.075	
	(0.114)	
	[0.501]	
Feedback	0.275**	
	(0.124)	
	[0.159]	
Male	0.070	0.087
	(0.117)	(0.120)
	[0.647]	[0.463]
Private x Male	–0.488***	
	(0.160)	
	[0.014]	
Feedback x Male	–0.203	
	(0.156)	
	[0.392]	
Ex post		0.080
		(0.118)
		[0.490]
Ex post x Male		–0.183
		(0.169)
		[0.218]
Risk Aversion	–0.006	0.001
	(0.012)	(0.020)
	[0.662]	[0.938]
Ability	–0.027***	–0.029***
	(0.005)	(0.006)
	[0.001]	[0.007]
Cognitive Ability	–0.000	–0.001
	(0.007)	(0.009)
	[0.9876]	[0.925]
Constant	3.712***	3.823***
	(0.261)	(0.334)
R^2	0.27	0.19
N	250	173

Notes: OLS regression with the weighted average-output quartiles as the dependent variable. The baseline group are women in the *Audience* treatment. In column (1), all 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender and, are considered. In column (2) only subjects in the *Audience* or the *Ex Post* treatment are considered, excluding one subject with technical problems. Numbers in parentheses indicate robust standard errors. Numbers in brackets show p-values based on wild cluster bootstrap t-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Appendix C. Heterogeneous effects based on performance

While the experiment was not designed to test for heterogeneous treatment effects based on performances, it is still interesting to investigate who, i.e., subjects who are close to the average or subjects who are far behind, are affected more strongly by the treatment variations. As the number of observations is too small for an extensive subgroup analysis, I relax the definition of overplacement instead in Table C.1. While in column (1) all subjects who state that they are better than average but whose performance is below the average performance in their session are classified as overplacing, the cut-off is lowered to include only subjects in the bottom 40% in column (2) or only the bottom 30% in column (3), respectively. I use OLS regressions and demean all continuous control variables for a more direct interpretation but the results are robust to using a logit model and non-standardized variables instead.

While it is a mechanical effect that the baseline level of confidence decreases as we lower the cut-off, for example the share of overplacing subjects in the *Audience* treatment decreases from 36% in column (1) to 16% in column (3), also the size of the treatment effect (see variables *Private* and *Feedback*) decreases when decreasing the cut-off (from left to right). This decrease indicates that it is mainly subjects close to the average performance who change their self-assessment.

Table C.1
Treatment effects for different cut-offs for overplacement.

	Overplacement = Yes		
	(1) below average	(2) bottom 40%	(3) bottom 30%
Private	–0.175*	–0.115	–0.104
	(0.094)	(0.082)	(0.069)
	[0.191]	[0.255]	[0.075]
Feedback	–0.253***	–0.158*	–0.109
	(0.092)	(0.082)	(0.071)
	[0.043]	[0.087]	[0.034]
Male	–0.094	–0.094	–0.066
	(0.103)	(0.088)	(0.076)
	[0.199]	[0.120]	[0.396]
Private x Male	0.126	0.103	0.082
	(0.135)	(0.113)	(0.093)
	[0.281]	[0.379]	[0.346]
Feedback x Male	0.212	0.112	0.060
	(0.132)	(0.109)	(0.092)
	[0.156]	[0.303]	[0.401]
Risk Aversion	–0.005	0.000	0.004
	(0.012)	(0.011)	(0.010)
	[0.690]	[0.967]	[0.783]
Cognitive Ability	–0.000	–0.000	–0.002
	(0.006)	(0.005)	(0.004)
	[0.968]	[0.935]	[0.625]
Constant	0.357***	0.238***	0.164***
	(0.075)	(0.068)	(0.059)
R^2	0.03	0.02	0.02
N	250	250	250

Notes: OLS regressions replicating the results of Table A.3 for different cut-offs to define overplacement. The baseline group are females in either the *Audience* treatment. All 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. Column (1) defines subjects as overplacing if they believe that they are better than the average but there performance is below the average of their session. Column (2) defines subjects as overplacing if they believe that they are better than the average but there performance is among the bottom 40% of their session. Column (3) defines subjects as overplacing if their performance is among the bottom 30%. All numerical control variables are demeaned for easier interpretability. Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on wild cluster bootstrapped t-tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Appendix D. What about underplacement?

In the main part of the paper, I focus on subjects who overplace because one aspect I am interested in is the motivational effect of overplacement. However, it is also interesting to investigate whether public observability and timing of self-assessments affect underplacement.

Fig. D.1 extends Figs. 1 and 2 by also showing the share of subjects who underplace (light gray bars with black outline) and the share of subjects who correctly guess their placement (gray bars). The pattern for underplacing subjects mirrors the effect for overplacing subjects, i.e., underplacement is higher when the self-assessment is private rather than public and when feedback is public. However, the increase is only statistically significant in the *Feedback* treatment ($p = 0.033$, see Table D.1).

In the lower panel of Fig. D.1 and in columns (2) and (3) of Table D.1, I investigate treatment effects by gender and observe that for women underplacement increases significantly from 2% in *Audience* to 16% in *Private* ($p = 0.014$). I find a similar effect for the introduction of public feedback (2% in *Audience* vs. 16% in *Feedback*, $p = 0.008$). In comparison, 12% of the male participants underplace in *Audience* and this number is not significantly different to the 8% in *Private* ($p = 0448$). Introducing public feedback does not significantly increase the share of men who underplace (11% in *Audience* vs. 15% *Feedback*, $p = 0.495$).

To test for gender differences in the reaction to the different treatments, I use a logistic regression in Table D.2 with interaction terms for gender and treatment. While the explanatory power of the regression is low, the coefficients point toward the expected direction, i.e., the odds to be underconfident when the self-assessment is private rather than public decrease less for men than for women ($p = 0.009$, see *Private x Male*). The same holds considering the effect of public feedback. The odds that a men is underconfident increase only 0.177 times as much as the odds that a women is underconfident when public feedback is introduced ($p = 0.014$, see *Feedback x Male*).

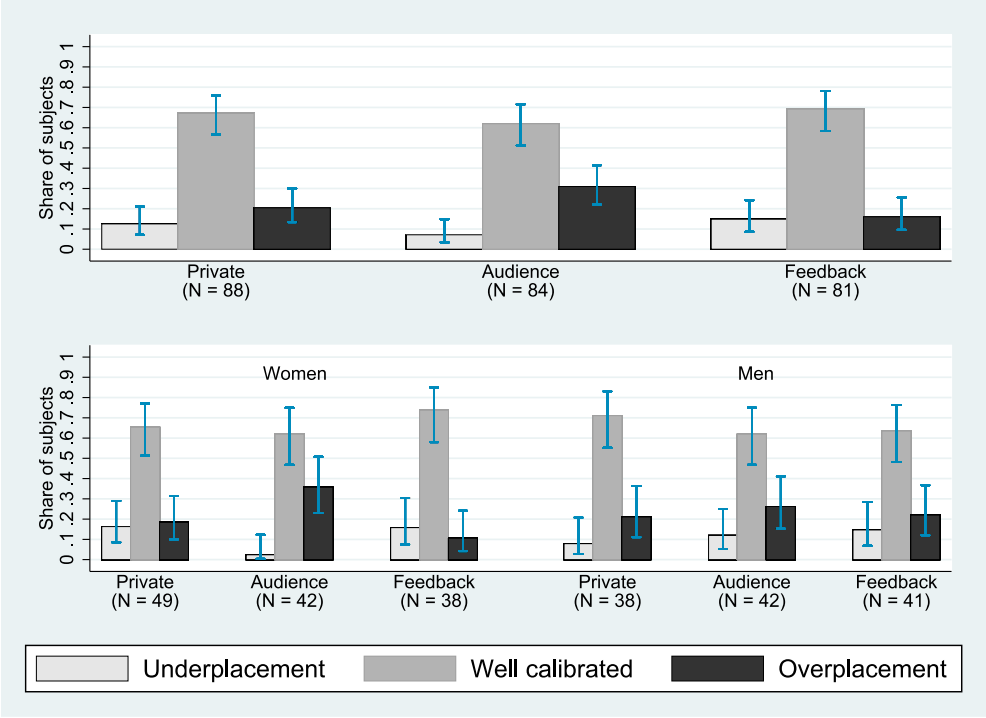


Fig. D.1. Treatment effects are driven by women. *Notes:* The figure shows the share of subjects who underplace (light gray bars with black outline), who correctly guess their placement (gray bars), and the share of subjects who overplace themselves (black bars) by treatment overall (upper panel) and split by gender (lower panel). Whiskers show Wilson confidence intervals (Brown et al., 2001). Overplacement is measured using the output in phase 1 as a proxy for actual ability. The figure includes the 253 subjects in the ex-ante treatments. Three subjects did not indicate their gender and are not included in the lower panel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table D.1
Underplacement by gender and treatment.

	Logistic regression (Odds Ratios)		
	Underplacement = Yes		
	(1) Pooled	(2) Women	(3) Men
Private	1.857	8.000*	0.634
	(0.991)	(8.701)	(0.489)
	[0.309]	[0.014]	[0.448]
Feedback	2.261	7.687*	1.269
	(1.193)	(8.532)	(0.828)
	[0.033]	[0.008]	[0.495]
Constant	0.077***	0.024***	0.135***
	(0.033)	(0.025)	(0.065)
Pseudo R^2	0.01	0.07	0.01
	OLS regression		
Private	0.054	0.139**	-0.040
	(0.045)	(0.058)	(0.067)
	[0.340]	[0.013]	[0.525]
Feedback	0.077	0.134**	0.027
	(0.049)	(0.064)	(0.075)
	[0.039]	[0.015]	[0.549]
Constant	0.071**	0.024	0.119**
	(0.028)	(0.024)	(0.051)
R^2	0.01	0.04	0.01
N	253	129	121

Notes: The upper panel shows odds ratios from logistic regressions and the lower panel shows an OLS regression.The baseline is the *Audience* treatment. All 253 subjects in the ex-ante treatments are considered in columns (1) while the sample is split by gender in the other columns. Three subjects did not indicate their gender and are therefore excluded in the respective columns.Numbers in parentheses indicate robust standard errors. Numbers in brackets show *p*-values based on cluster bootstrapped score tests (999 replications) in the upper sample and wild cluster bootstrap *t*-tests (999 replications) in the lower panel to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table D.2
Underplacement varies across treatments.

	Underplacement = Yes	
	(1)	(2)
Private	8.000*	8.050*
	(8.685)	(8.816)
	[0.015]	[0.014]
Feedback	7.687*	7.320*
	(8.516)	(8.279)
	[0.006]	[0.003]
Male	5.541	5.027
	(6.210)	(5.702)
	[0.110]	[0.037]
Private x Male	0.079*	0.080*
	(0.105)	(0.109)
	[0.011]	[0.009]
Feedback x Male	0.165	0.177
	(0.212)	(0.232)
	[0.017]	[0.014]
Risk Aversion		1.036
		(0.087)
		[0.777]
Cognitive Ability		1.036
		(0.054)
		[0.514]
Constant	0.024***	0.010***
	(0.025)	(0.015)
Pseudo R^2	0.04	0.05
N	250	250

Notes: Coefficients are reported as odds ratios. The baseline group are women in the *Audience* treatment. All 253 subjects in the ex-ante treatments, except three subjects who did not indicate their gender, are considered. Underplacement is based on output in phase 1. Numbers in parentheses indicate robust standard errors. Numbers in brackets show p-values based on cluster bootstrapped score tests (999 replications) to account for potential correlation of standard errors within sessions while taking the small number of clusters into account. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Appendix E. Comparison to previous results

To put the results of this paper into perspective, I investigate treatment differences in overconfidence for papers which also manipulate social image concerns of self assessments on an objective performance scale. Based on my literature review there are three papers fulfilling this requirement, namely Ewers and Zimmermann (2015), short EZ, Schwardmann and van der Weele (2019), short SV, and Ludwig et al. (2017), short LFT.[36]

The data of all three papers is publicly available. As the focus of these papers is not to identify gender differences in image concerns, I conducted my own calculations to identify gender differences in self-assessments. For better interpretability of effect sizes I define a binary overplacement variable for each of the studies and investigate how the share of overplacing subjects is affected by treatment. Based on the structure of the experiments and the data, subjects are defined as overplacing if they believe that they are better than the average (median) but their performance was worse than the average (median) for this paper and EZ (for SV and LFT).

Table E.1 shows that the result that subjects' self-assessments increase when they are made publicly (compare *Private* and *Audience*) is robust across the different experimental designs and subject groups. The drop in self-assessments if performance is observable as well (compare *Audience* and *Feedback*) is less robust due to gender differences in the reaction to public feedback. While women's confidence drops in the light of public feedback (ranging between -19pp in LFT and -25pp in this paper), the effects on men's confidence are less clear (ranging from -13pp in EZ and +3pp in LFT).

[36] Another potential candidate for a comparison is Buser et al. (2021). However, it is not included in this comparison because subjects publicly announce their willingness to enter a competition rather than their self-assessments and self-assessments are elicited after participants learn whether or not they have won the tournament which very likely affects their self-assessment and reduce comparability.

Table E.1
Treatment effects in the literature.

Share overplacing	Women			Man		
	Private	Audience	Feedb.	Private	Audience	Feedb.
H (M = 120, W = 129)	18%	36%	11%	21%	26%	21%
EZ (M = 100, W = 111)	22%	32%	8%	21%	41%	28%
SV (M = 113, W = 175)	24%	28%		21%	25%	
SV^r (M = 165, W = 235)	24%	29%		22%	31%	
LFT (M = 51, W = 49)		36%	17%		20%	23%

Notes: H: Overplacement as defined in the main part of the paper. EZ: Data from Ewers and Zimmermann (2015). Subjects are defined as overplacing if they stated that there are better than average but their performance was worse than the average as defined by the authors. SV: Data from Schwardmann and van der Weele (2019). Subjects are defined as overplacing if their prior to be among the top two performer (in their groups of four) is greater than 50% but they were not among the top two of their group. SV^r: is based on the data of the replication data presented in Schwardmann and van der Weele (2019). Overplacement is defined as explained above. LFT: Data from Ludwig et al. (2017). Subjects are defined as overplaced if they believe their rank is better than 11 (out of 22) but their actual rank in their session is worse.

Appendix F. Instructions

Instructions for participants

(Feedback treatment)

General explanation

You are now taking part in an economic experiment. During the experiment you can earn money. Therefore it is important that you read the following instructions carefully.

These instructions are solely for your private information. Do not communicate with other participants during the experiment! If you have any questions, please raise your hand and an assistant will help you.

During the experiment your earnings will be calculated in points. Your earnings from the experiment will be paid to you in cash right after the experiment. Your earnings in points will be converted into cash according to the following exchange rate:

13Points = 1EUR

The experiment consists of several phases.

Explanation for the 1^{st} phase

You will be asked to make ten decisions between two options A and B, as shown in the example below.

	Option A	Option B	Your Choice
Decision 1:	Heads: You get 5 points. Tails:You get 3 points.	Heads: You get 7 points. Tails:You get 1 point.	○ A ○ B

You will only get paid for one of your decisions. At the end of the experiment, we will ask one of the participants to draw a number from 1 to 10, to determine which of the decisions will be paid out. Furthermore, we will toss a coin publicly to get Heads or Tails. This means that Heads and Tails are equally likely.

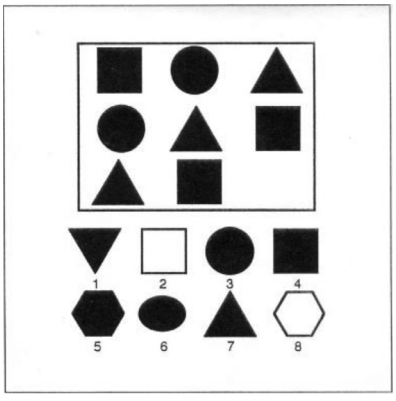
For example: Assume decision 1 has been picked to determine your payoff and you decided on option A. Then you have a 50% chance that the coin shows Heads and you receive 5 points and a 50% chance that it will show Tails and you receive 3 points.

Explanation for the 2^{nd} phase

You will see 36 screens in sequence. On each screen you will find a task. We give you 20 minutes and we ask you to solve as many tasks as possible correctly. For each **correctly** solved task you receive **2 points**. You will not get feedback on the accuracy of your answers before the end of the experiment.

Example of a task:

Here is an example of a task. The correct solution of this task is symbol number 3. There is a square, a circle, and a triangle in each row. In the third row, there is a triangle and a square, but no circle. Therefore, symbol number 3 is the correct answer.



The experiment will start soon. Please raise your hand if you have any questions.

Explanation for the 3^{rd} phase

What to do in the 3rd phase:

In this phase you will calculate cross-sums. You have to sum up the sequence of digits. Here is an example:

5 7 8 0 3

Your task is to calculate the cross-sum, that is: $5+7+8+0+3$

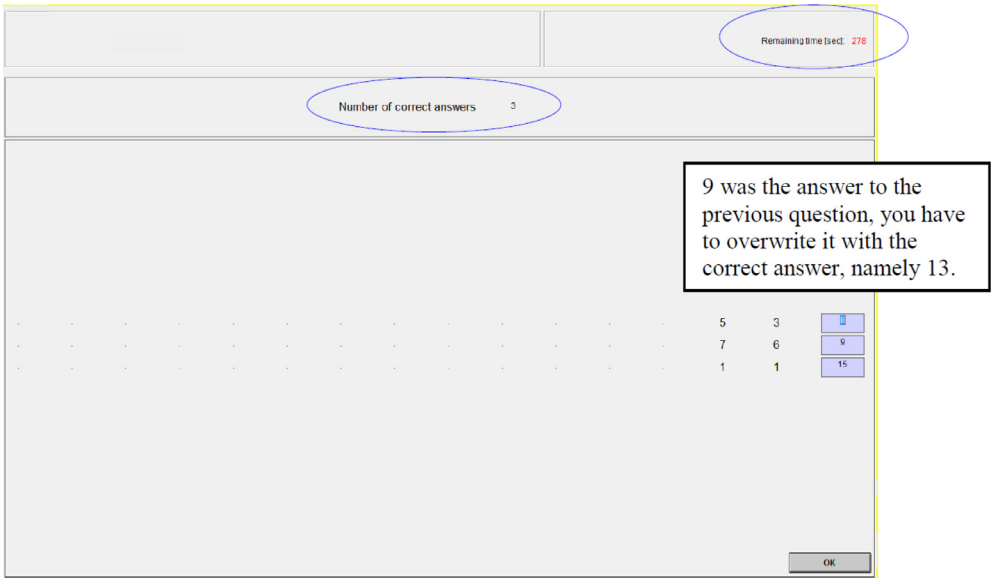
The correct answer in this example is 23.

How your earnings are calculated:

The 3rd phase lasts for 5 minutes (300 seconds). For each correct cross-sum you receive a payment of **0.7 points**.

You solve cross-sums on screens (see example below). Each screen contains 3 cross-sums. Only if you solve all three calculations correctly and click the "OK"-button, the next screen will be shown. If you make a mistake in one of the calculations, the program tells you where you made the mistake and you have to revise your answer (This will be displayed at the bottom left corner).

On the top of the screen you see the **remaining time** to make your calculations and the **number of points earned** so far. Note that as you press "OK" and start a new set of cross sums, the answers to the previous cross sums remain.



Warning: when you enter a new screen of cross-sums the answers to the cross-sums from the previous screen of cross-sums will still be in the **answering fields** (like on the screen); these have to be **overwritten**. You can use the mouse or the TAB-button to maneuver from one cross-sum to the next. However, for many people the easiest way to maneuver from one crosssum to the next cross-sum on the screen is by using the TAB-button, only using the mouse to click the "OK"-button.

In the beginning the cross-sums consist of 2 digits. After every 5 screens one digit is added to each cross-sum on the screen; i.e. the 3 cross-sums from the 6^{th} screen each consist of 3 digits, from the 11^{th} screen of 4 digits, from the 16^{th} screen of 5 digits, and so on.

At the end of phase 3 you will be informed about how many cross-sums you have solved and your earnings in this particular phase.

Training phase:

Before entering phase 3, we conduct a 4 minute training phase to familiarize you with calculating cross-sums and entering your decisions on the computer. Compared with phase 3 there are the following two differences:

1. The difficulty of the cross-sums increases after every two screens of cross-sums (6 cross-sums) instead of after every five screens. This way you can experience the difference in difficulty of cross-sums with different numbers of digits.
2. You will not earn any money in the training phase. However, previous studies have shown that people become considerably faster at doing cross-sums when being familiar with them (and thus increase their potential earnings of the future phases of the experiment) and we therefore strongly encourage you to do so.

The experiment will start soon. Please raise your hand if you have any questions.

Explanation for the 4^{th} phase

What to do in the 4^{th} phase:

In the 4th phase, you have to solve the same kind of tasks as in phase 3, but an additional task, the SWITCH-Task is introduced and your payment is calculated differently. In the SWITCH- Task you don't have to calculate cross-sums anymore.

How your earnings are calculated:

The 4^{th} phase lasts for 20 minutes (1200 seconds). We will now explain how your earnings are determined.

1. Your earnings from the solution of tasks: for each correct cross-sum you receive **0.7 points**.
2. Your earnings from the SWITCH-Task: during phase 4 you can switch to the so called SWITCH-Task at any time. In the remaining time after switching to the SWITCH-Task until the end of the block you earn **1 point for each 15 seconds**.

Example for earnings from the SWITCH-Task:

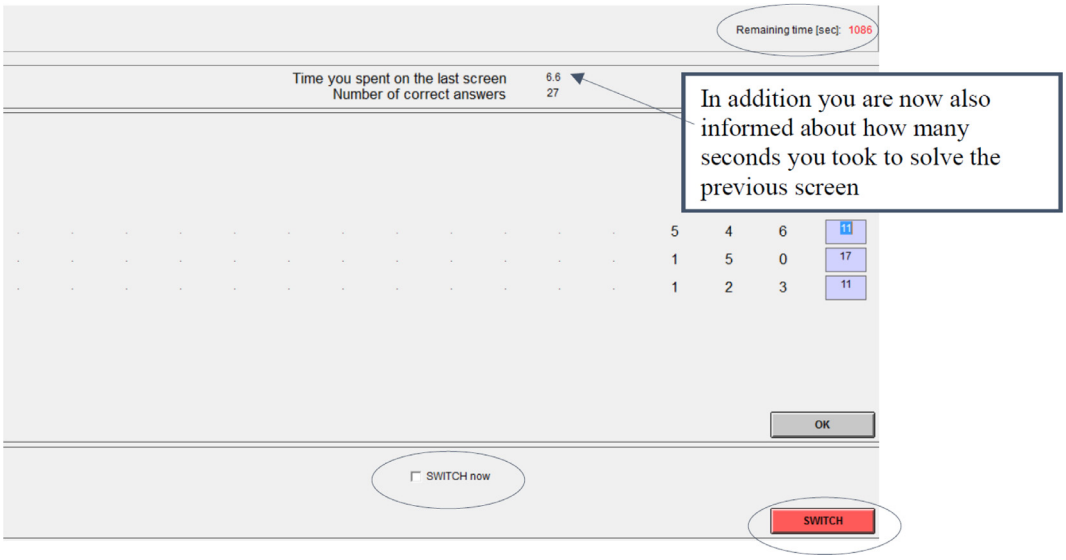
The phase lasts for 1200 seconds. Suppose you switch to the SWITCH-Task after 800 seconds. This means that 400 seconds remain till the end of the round. For these 400 seconds you earn **26.6 (400/15) points**.

After the 20 minutes are over, you will be informed about the number of tasks you have solved, as well as your earnings from this phase.

On the next page, we will show you what the screens look like.

What the screens look like:

On the top of the screen you see the **remaining time** to make your calculations and the **number of points earned** so far. Note that as you press "OK" and start a new set of cross-sums, the answers to the previous cross-sums will still be in the **answering fields** (like on the screen); these have to be **overwritten**.



SWITCH-Task:

If you press the **"SWITCH"**-button on any of the screens, you switch to the SWITCH task. Remember that you earn **1 point** for each **15 seconds** you have spent on the SWITCH task. To ensure that you do not switch to the SWITCH task by accident you have to mark the **"SWITCH" checkbox** before you can press the "SWITCH"-button.

Important: Once you have switched to the SWITCH task you cannot switch back to doing cross-sums.

Once you have switched to the SWITCH task, you will wait until the 20 minutes are up. If you are in the SWITCH task, you must remain seated and you are not allowed to communicate with others.

Important: As you already know, each screen consists of 3 cross-sums, each worth 0.7 points.

Therefore you can earn **2.1 points** for each screen. In the SWITCH-Task, you earn **1 point for each 15 seconds**.

This means: if you are fast in solving cross-sums, you earn more in the cross-sum task than in the SWITCH-Task. On the contrary, if you are slow in solving the cross-sums, you earn less than in the SWITCH-Task. Concretely, you receive higher earnings in the SWITCH-Task, if it takes you more than **2.1* 15 = 32 seconds** (rounded) to solve the 3 cross-sums on one screen.

The experiment will start soon. Please raise your hand if you have any questions.

Explanation for the 5^{th} phase

What to do in the 5^{th} phase:

The structure of phase 5 is the same as in phase 4. It lasts for 20 minutes and you earn a piece rate for solving cross sums (2.1 points per screen) and have the option to switch to the SWITCH task. In the SWITCH task you do not have to work and earn 1 point every 15 seconds.

This means: if you are fast in solving cross-sums, you earn more in the cross-sum task than in the SWITCH-Task. On the contrary, if you are slow in solving the cross-sums, you earn less than in the SWITCH-Task. Concretely, you receive higher earnings in the SWITCH-Task, if it takes you more than **2.1*15 = 32 seconds** (rounded) to solve the 3 cross-sums on one screen. However, this time we ask you to state how good you think you are at solving cross sums compared to a similar group of students, who did a similar experiment some time ago.

The self-assessment consists of three steps:

Step 1: We will ask you to do a **private self-assessment** as explained on the next page.

Step 2: We will ask you to state whether you think that you are better or worse than the average participant of the other group.

Step 3: After all participants entered their assessments into the computer, all participants must report their assessment from **step 2** to the other participants in this session. You will be called up individually one after the other. Once it is your turn, you have to stand up, say your name and report your assessment.

So if you stated that you think your performance in phase 5 will be better than the average of the other group, then you have to stand up after you were called and say: "My name is … and I think I am better than the average of the other group."

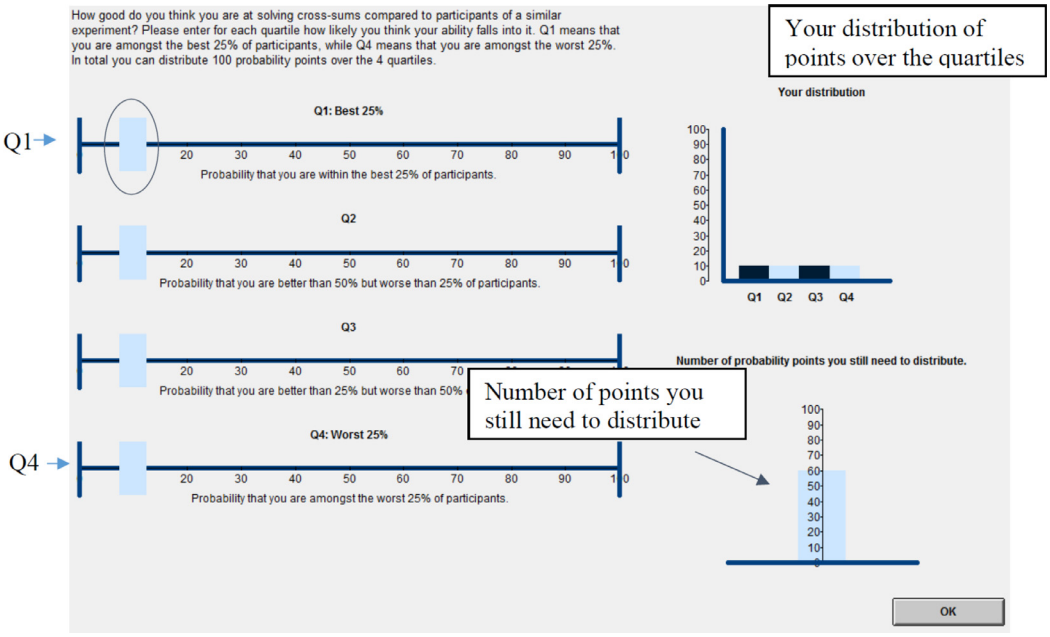
So if you stated that you think your performance in phase 5 will be worse than the average of the other group, then you have to stand up after you were called and say: "My name is … and I think I am worse than the average of the other group."

Private Self-assessment:

Before Phase 5 starts, we kindly ask you to think about how good you think you are at solving cross-sums compared to a group who participated in an earlier version of this experiment (in which participants were not asked to do the self-assessment). To be more precise, we would like to know how likely you think it is that you will end up in each of the five quartiles presented in the figure below.

The first quartile **(Q1)** means that you are among the 25% who are best at solving cross-sums. The forth **(Q4)** means that you are among the 25% who are worst at solving the cross-sums. In total you can distribute **100** probability points over the four quartiles. In the bottom right of the screen

you see how many points you have used already and in the upper right you see how you have distributed the points. To distribute points you have to click on the **light blue bars**, hold the mouse button down and drag the bar to the number you would like to distribute.



In the example above, the participant has so far given 10 points to each of the quartiles and still has to distribute 60 points. If you have reached the limit of points you can distribute, a message will pop up. Once there are no points to distribute left, and you are happy with your choice, please press the “OK”button.

After the 20 minutes:

At the end of Phase 5, we will inform you about the number of cross-sums you have solved in this phase. We will also use your performance in this phase to see if your self-assessment was correct.

This means that we will compare the number of cross-sums you have solved in phase 5 to the performance of a similar group of students, who did a similar experiment some time ago.

After showing you a general payoff screen, on which you can see how much you have earned in each phase of the experiment, we will show you again your self-assessment from step 1 and mark your actual quartile in red.

Additionally, you will have to stand up again and we will inform you publicly whether you have been better or worse than the average of the other group. This means that the other participants will know if what you said in Step 3 was correct or not.

The experiment will start soon. Please raise your hand if you have any questions.

References

Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. Am Econ Rev 101 (2), 470–492.

Bénabou, R., Tirole, J., 2002. Self-confidence and personal motivation. Q J Econ 117 (3), 871–915.

Bengtsson, C., Persson, M., Willenhag, P., 2005. Gender and overconfidence. Econ Lett 86 (2), 199–203. https://EconPapers.repec.org/RePEc:eee:ecolet:v:86:y:2005:i:2:p:199-203

Benoît, J.-P., Dubra, J., Romagnoli, G.,. Belief elicitation when more than money matters: controlling for "control".American Economic Journal: Microeconomics. forthcoming

Berger, J., Harbring, C., Sliwka, D., 2013. Performance appraisals and the impact of forced distribution-an experimental investigation. Manage Sci 59 (1), 54–68.

Blumkin, T., Ruffle, B., Ganun, Y., 2010. Are Income and Consumption Taxes Ever Really Equivalent? Evidence from a Real-Effort Experiment with Real Goods. IZA Discussion Paper. Institute for the Study of Labor (IZA). No. 5145

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A., 2019. Beliefs about gender. Am Econ Rev 109 (3), 739–773. doi:10.1257/aer.20170007.

Borghans, L., Golsteyn, B.H.H., Heckman, J.J., Meijers, H., 2009. Gender differences in risk aversion and ambiguity aversion. J Eur Econ Assoc 7 (2/3), 649–658. http://www.jstor.org/stable/40282781

Brandts, J., Gërxhani, K., Schram, A., 2020. Are there gender differences in status-ranking aversion? J Behav Exp Econ 84, 101485. doi:10.1016/j.socec.2019.101485. https://www.sciencedirect.com/science/article/pii/S2214804319301612

Brown, L.D., Cai, T.T., DasGupta, A., 2001. Interval estimation for a binomial proportion. Stat Sci 16 (2), 101–117. http://www.jstor.org/stable/2676784

Burks, S.V., Carpenter, J.P., Goette, L., Rustichini, A., 2013. Overconfidence and social signalling. Rev Econ Stud 80 (3 (284)), 949–983. http://www.jstor.org/stable/43551452

Bursztyn, L., Egorov, G., Jensen, R., 2018. Cool to be smart or smart to be cool? understanding peer pressure in education. Rev Econ Stud 86 (4), 1487–1526.

Buser, T., Niederle, M., Oosterbeek, H., 2014. Gender, competitiveness, and career choices. Q J Econ 129 (3), 1409–1447. doi:10.1093/qje/qju009.

Buser, T., Peter, N., Wolter, S.C., 2017. Gender, competitiveness, and study choices in high school: evidence from switzerland. Am Econ Rev 107 (5), 125–130. doi:10.1257/aer.p20171017. http://www.aeaweb.org/articles?id=10.1257/aer.p20171017

Buser, T., Ranehill, E., van Veldhuizen, R., 2021. Gender differences in willingness to compete: the role of public observability. J Econ Psychol 83, 102366. doi:10.1016/j.joep.2021.102366. https://www.sciencedirect.com/science/article/pii/S0167487021000064

Cameron, A.C., Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. Journal of Human Resources 50 (2), 317–372.

Charness, G., Gneezy, U., Rasocha, V., 2021. Experimental methods: eliciting beliefs. J Econ Behav Organ 189, 234–256. doi:10.1016/j.jebo.2021.06.032. https://www.sciencedirect.com/science/article/pii/S0167268121002717

Chen, S., Schildberg-Hörisch, H., 2019. Looking at the bright side: the motivation value of overconfidence. Eur Econ Rev 120, 103302.

Clark, D., Gill, D., Prowse, V., Rush, M., 2017. Using Goals to Motivate College Students: Theory and Evidence from Field Experiments. NBER Working Papers. National Bureau of Economic Research, Inc. No. 23638

Coffman, K.B., 2014. Evidence on self-stereotyping and the contribution of ideas. Q J Econ 129 (4), 1625–1660. doi:10.1093/qje/qju023.

Croson, R., Gneezy, U., 2009. Gender differences in preferences. J Econ Lit 47 (2), 448–474. doi:10.1257/jel.47.2.448.

Dohmen, T., Falk, A., 2011. Performance pay and multidimensional sorting: productivity, preferences, and gender. Am Econ Rev 101 (2), 556–590. doi:10.1257/aer.101.2.556. http://www.aeaweb.org/articles?id=10.1257/aer.101.2.556

Dreber, A., von Essen, E., Ranehill, E., 2014. Gender and competition in adolescence: task matters. Exp Econ 17 (1), 154–172. doi:10.1007/s10683-013-9361-0.

Drichoutis, A.C., Lusk, J.L., 2016. What can multiple price lists really tell us about risk preferences? J Risk Uncertain 53 (2–3), 89–106. doi:10.1007/s11166-016-9248-5.

Eckartz, K.M., 2014. Task Enjoyment and Opportunity Costs in the Lab - The Effect of Financial Incentives on Performance in Real Effort Tasks. Jena Economic Research Papers, No. 2014-005. Max-Planck-Institut für Ökonomik und Universität Jena. https://ideas.repec.org/p/jrp/jrpwrp/2014-005.html

Eil, D., Rao, J.M., 2014. The good news-bad news effect: asymmetric processing of objective information about yourself. Am Econ J: Microecon 3 (2), 114–138. doi:10.1257/mic.3.2.114.

Erkal, N., Gangadharan, L., Boon, H.K., 2018. Monetary and non-monetary incentives in real-effort tournaments. Eur Econ Rev 101, 528–545.

Ewers, M., Zimmermann, F., 2015. Image and misreporting. J Eur Econ Assoc 13 (2), 363–380. doi:10.1111/jeea.12128. https://onlinelibrary.wiley.com/doi/abs/10.1111/jeea.12128

Exley, C.L., Kessler, J.B., 2019. The Gender Gap in Self-Promotion. NBER Working Papers, No. 26345. National Bureau of Economic Research, Inc. http://www.nber.org/papers/w26345

Fischbacher, U., 2007. Z-tree: zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2), 171–178.

Gächter, S., Huang, L.B., Sefton, M., 2016. Combining "real effort" with induced effort costs: the ball-catching task. Exp Econ 19 (4), 687–712. doi:10.1007/s10683-015-9465-9.

Gneezy, U., Leonard, K.L., List, J.A., 2009. Gender differences in competition: evidence from a matrilineal and a patriarchal society. Econometrica 77 (5), 1637–1664. doi:10.3982/ECTA6690. http://doi.wiley.com/10.3982/ECTA6690

Gneezy, U., Niederle, M., Rustichini, A., 2003. Performance in competitive environments: gender differences. Q J Econ 118 (3), 1049–1074. doi:10.1162/00335530360698496.

Gneezy, U., Saccardo, S., Serra-Garcia, M., van Veldhuizen, R., 2020. Bribing the self. Games Econ Behav 120, 311–324. doi:10.1016/j.geb.2019.12.010. https://www.sciencedirect.com/science/article/pii/S0899825619301939

Goerg, S., Kube, S., 2012. Goals (th)at Work - Goals, Monetary Incentives, and Workers' Performance. Discussion Paper Series of the Max Planck Institute for Research on Collective Goods, No. 2012 19. Max Planck Institute for Research on Collective Goods. https://EconPapers.repec.org/RePEc:mpg:wpaper:2012_19

Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with orsee. Journal of the Economic Science Association 1 (1), 114–125. doi:10.1007/s40881-015-0004-4.

Grossman, Z., Owens, D., 2012. An unlucky feeling: overconfidence and noisy feedback. J Econ Behav Organ 84 (2), 510–524. doi:10.1016/j.jebo.2012.08.006.

Haeckl, S., Sausgruber, R., Tyran, J.-R., 2018. Work Motivation and Teams. Discussion Papers, No. 18-08. University of Copenhagen. Department of Economics. https://EconPapers.repec.org/RePEc:kud:kuiedp:1808

Hardies, K., Breesch, D., Branson, J., 2013. Gender differences in overconfidence and risk taking: do self-selection and socialization matter? Econ Lett 118 (3), 442–444. doi:10.1016/j.econlet.2012.12.004. http://www.sciencedirect.com/science/article/pii/S0165176512006404

Hayashi, A.T., Nakamura, B.K., Gamage, D., 2013. Experimental evidence of tax salience and the labor-leisure decision: anchoring, tax aversion, or complexity? Public Finance Rev. 41 (2), 203–226. doi:10.1177/1091142112460726.

Hill, S.J., 2020. Following through on an intention to vote: present bias and turnout. Political Science Research and Methods 8, 803–810.

von Hippel, W., Trivers, R., 2011. The evolution and psychology of self-deception. Behav Brain Sci 34 (1), 1–16. doi:10.1017/S0140525X10001354.

Koch, A.K., Nafziger, J., 2011. Self-regulation through goal setting. Scand J Econ 113 (1), 212–227. doi:10.1111/j.1467-9442.2010.01641.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9442.2010.01641.x

Koch, A.K., Nafziger, J., 2020. Motivational goal bracketing: an experiment. J Econ Theory 185, 104949.

Köszegi, B., 2006. Ego utility, overconfidence, and task choice. J Eur Econ Assoc 4, 673–707. doi:10.1162/JEEA.2006.4.4.673.

Köszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. Q J Econ 121 (4), 1133–1165. doi:10.1093/qje/121.4.1133.

Köszegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. Am Econ Rev 97 (4), 1047–1073.

Krawczyk, M., 2019. What should be regarded as deception in experimental economics? evidence from a survey of researchers and subjects. J Behav Exp Econ 79, 110–118. doi:10.1016/j.socec.2019.01.008. https://www.sciencedirect.com/science/article/pii/S221480431830329X

Latham, G.P., Locke, E.A., 1991. Self-regulation through goal setting. Organ Behav Hum Decis Process 50 (2), 212–247. doi:10.1016/0749-5978(91)90021-K. Theories of Cognitive Self-Regulation, https://www.sciencedirect.com/science/article/pii/074959789190021K

van Lent, M., Souverijn, M., 2017. Goal Setting and Raising the Bar: A Field Experiment. Tinbergen Institute Discussion Paper, No. 17-001/VII. Tinbergen Institute.

Locke, E.A., Latham, G.P., 1990. A Theory of Goal Setting & Task Performance. Prentice-Hall, Inc.

Ludwig, S., Fellner-Röhling, G., Thoma, C., 2017. Do women have more shame than men? an experiment on self-assessment and the shame of overestimating oneself. Eur Econ Rev 92, 31–46. doi:10.1016/j.euroecorev.2016.11.007. http://www.sciencedirect.com/science/article/pii/S0014292116302197

Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S., 2011. Managing Self-Confidence: Theory and Experimental Evidence. NBER Working Papers, No. 17014. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/17014.html

Mohnen, A., Pokorny, K., Sliwka, D., 2008. Transparency, inequity aversion, and the dynamics of peer pressure in teams: theory and evidence. J Labor Econ 26 (4), 693–720. doi:10.1086/591116.

Moore, D.A., Healy, P.J., 2008. The trouble with overconfidence. Psychol Rev 115 (2), 502–517. doi:10.1037/0033-295X.115.2.502.

Nekby, L., Thoursie, P.S., Vahtrik, L., 2008. Gender and self-selection into a competitive environment: are women more overconfident than men? Econ Lett 100 (3), 405–407. doi:10.1016/j.econlet.2008.03.005. http://www.sciencedirect.com/science/article/pii/S0165176508000888

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? do men compete too much? Q J Econ 122 (3), 1067–1101. doi:10.1162/qjec.122.3.1067.

Niederle, M., Vesterlund, L., 2008. Gender differences in competition. Negotiation Journal 24 (4), 447–463. doi:10.1111/j.1571-9979.2008.00197.x.

Reuben, E., Sapienza, P., Zingales, L., 2014. How stereotypes impair women's careers in science. Proc Natl Acad Sci 111 (12), 4403–4408. doi:10.1073/pnas.1314788111. https://www.pnas.org/content/pnas/111/12/4403.full.pdf

Ring, P., Neyse, L., David-Barett, T., Schmidt, U., 2016. Gender differences in performance predictions: evidence from the cognitive reflection test. Front Psychol 7, 1680. doi:10.3389/fpsyg.2016.01680. https://www.frontiersin.org/article/10.3389/fpsyg.2016.01680

Robinson, M.D., Ryff, C.D., 1999. The role of self-deception in perceptions of past, present, and future happiness. Pers Soc Psychol Bull 25 (5), 596–608.

Roodman, D., Nielsen, M.Ø., MacKinnon, J.G., Webb, M.D., 2019. Fast and wild: bootstrap inference in stata using boottest. Stata J 19 (1), 4–60.

Schlag, K., Tremewan, J., 2021. Simple belief elicitation: an experimental evaluation. J Risk Uncertain 62 (2), 137–155. doi:10.1007/s11166-021-09349-6.

Schwardmann, P., van der Weele, J., 2019. Deception and self-deception. Nat Hum Behav 3, 1055–1061. doi:10.1038/s41562-019-0666-7.

Smithers, S., 2015. Goals, motivation and gender. Econ Lett 131, 75–77. doi:10.1016/j.econlet.2015.03.030. https://www.sciencedirect.com/science/article/pii/S0165176515001317

Thoma, C., 2016. Under- Versus overconfidence: an experiment on how others perceive a biased self-assessment. Exp Econ 19 (1), 218–239. doi:10.1007/s10683-015-9435-2. https://ideas.repec.org/a/kap/expeco/v19y2016i1p218-239.html

Trautmann, S.T., van de Kuilen, G., 2015. Belief elicitation: a horse race among truth serums. Econ J 125 (589), 2116–2135. doi:10.1111/ecoj.12160. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12160

van Veldhuizen, R., 2018. Gender Differences in Tournament Choices: Risk Preferences, Overconfidence or Competitiveness? Rationality and Competition Discussion Paper Series. No. 14

Weber, M., Schram, A., 2017. The non-equivalence of labour market taxes: areal-effort experiment. Econ J 127 (604), 2187–2215. doi:10.1111/ecoj.12365. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12365