The Social Aspects of Job Competitions: Gender Gaps in Self-Assessment*

Christine Alamaa[1] ✉

[1]Swedish Institute for Social Research (SOFI), Stockholm University

Job Market Paper

November 20, 2025
[Latest Version]

Abstract

Competitions, particularly in professional contexts, convey information about participants' relative ability, yet they also reveal their self-beliefs. They transform personal judgements into common knowledge that others can observe and evaluate. This paper studies how exposing individuals' competitive information affects how they state their rank beliefs, and whether these effects differ by gender. I propose a framework in which individuals incur psychological costs when inaccuracies in their self-beliefs are revealed, either privately or publicly, and test its predictions in a repeated laboratory experiment ($N = 544$). Without feedback or observability of belief accuracy, women assess themselves more moderately than men, while introducing public observability reduces overconfidence for both genders. In contrast, private feedback generates gender-specific learning patterns: women adjust more after overestimation, men after underestimation. Together, the findings suggest that belief exposure as a distinct social mechanism in competition helps explain how gender differences in self-assessment emerge and persist in labour markets.

Keywords: Feedback and observability, self-assessment, gender differences, lab experiment, labour-market behaviour

JEL Codes: C91, D91, J16

*I am grateful for valuable comments and suggestions from Erik Lindqvist, Jenny Säve-Söderbergh, Johannes Hagen, Michèle Belot, Sule Alan, Chloé Le Coq, Louis-Pierre Lepage, Max Brès, Giulia Vattoune, Natalia Montinari, Emma von Essen, and other GAINS members, and participants at the IIES and SOFI Brown Bag Series, the NPPS Conference 2024, the SABE 2025 Annual Conference, and the IFAU Seminars 2025. I am also truly thankful to Tommaso Batistoni for invaluable coding assistance. Operative support in conducting the experiment was provided by the BLESS Lab at the University of Bologna. Financial support from the Institute for Evaluation of Labour Market and Education Policy (IFAU), the Swedish Research Council (VR), and FORTE–Swedish Research Council for Health, Working Life and Welfare is gratefully acknowledged. All remaining errors are my own.

Corresponding author: Christine Alamaa, christine.alamaa@sofi.su.se, Swedish Institute for Social Research.

1 Introduction

Competitions are integral to professional life. Yet, they do more than allocate positions and pay—they generate information, going beyond performance and advancement. For example, in hiring, promotion, and salary negotiations, individuals are compared and form beliefs about their abilities and relative standing, exposing potential inaccuracies in self-assessments to themselves and to others. Crucially, such situations may transform private judgements into common information. These social dynamics raise the question of how individuals process and respond to such experiences and whether some groups differ systematically in ways that influence career advancement. Although the role of monetary considerations in competitive environments is relatively well understood, much less is known about whether social considerations independently influence why some individuals persist while others withdraw.

Gender patterns in labour-market competitions illustrate where social dynamics may operate. An empirical literature shows that women and men sort into different positions and competitive settings, producing persistent gaps in outcomes and careers (Azmat and Petrongolo, 2014). With differences in education or productivity unable to account for these gaps (Blau and Kahn, 2017), explanations introduced psychological factors—risk preferences, "competitiveness", and self-confidence—yet evidence remains mixed (*see,* Van Veldhuizen, 2022). Recent evidence begins to uncover the mechanisms behind gendered self-sorting, documenting divergence in application behaviour: women apply less often to high-tier positions, even when similarly qualified (Fluchtmann et al., 2024).[1] This behavioural pattern aligns with evidence that even among top economists, women give less confident self-assessments than men (Sarsons and Xu, 2021). However, identifying decision-making in competition and application behaviour remains challenging, as most data capture participants but not those who refrain.

While existing research typically links gender gaps in competitive behaviour to stable psychological traits, this paper provides a different perspective. I propose that an inherent feature of competition—its revelation of comparative information to participants and others—might impose psychological costs tied to emotional or reputational risks. When individuals learn their self-assessments were inaccurate—"I claimed to be ranked well, and now both I and others know I was wrong"—competitions might trigger responses to exposing a misjudged self-evaluation.[2] Such behavioural responses may further differ depending on whether they arise from being publicly observable or from self-scrutiny—for example, whether the information is revealed to a recruiter or recognised by the candidate herself.[3] An additional question regards timing: do individuals adjust competitive behaviour and self-presentation in anticipation of such social costs, or only after exposure? Because social considerations are intertwined with material and strategic ones in the labour market, my setting offers an opportunity to examine their effects on self-assessment and whether they vary by gender.

To formalise these mechanisms, I develop a conceptual framework in which individuals decide how

[1] Fluchtmann et al. (2024) use Danish administrative data on job applications from unemployment-insurance records. Similar patterns persist among highly qualified individuals (Cortés et al., 2023) and in education, where women are less likely to sort into competitive tracks (Buser et al., 2014).

[2] Reactions in competition entry to the exposure of positional standing—termed status-ranking aversion by Brandts et al. (2020), who also study gender differences—are analysed in their work.

[3] Prior work considers both private and public scrutiny as contexts giving rise to socially framed, self-evaluative emotions even in private settings (*see* Bénabou and Tirole, 2006; Sedikides and Strube, 1997; Dweck, 1999). Such reactions are commonly classified as a type of *moral emotion*, known as self-conscious emotions—including shame, pride, or embarrassment—that engage the social self.

confidently to assert their production-based rank. In the framework, agents trade off expected material payoffs against the psychological costs of potentially revealing self-misjudgment. I model how agents form and report beliefs across environments that differ in whether, and by whom, their belief accuracy is observed. I then track how these beliefs evolve dynamically through successive interactions that reveal information in repeated rounds over time. The framework accommodates heterogeneity in agents' initial beliefs and in their responsiveness to feedback. Finally, it generates conceptual predictions about agents' belief updating, the effects of self- and social exposure, and group-level patterns in self-confidence over time.

I design a laboratory experiment that varies the visibility of agents' belief accuracy and tracks how they adjust their stated rank beliefs over time. The experiment involved 576 participants at the Bologna Laboratory for Experiments in Social Science (BLESS),[4] who were randomly assigned to the role of *agent* or *principal*, without strategic interaction between roles. Agents completed a performance task and reported a probabilistic self-assessment of their rank, with incentives tied to both task scores and the accuracy of these rank beliefs.[5] Participants were randomised into four conditions that varied whether belief accuracy was undisclosed (*Control*) or disclosed only to agents (*Private*), only to principals (*Public*), or to both (*Joint*). To distinguish anticipation effects (before misjudgment is revealed) from experience effects (after receiving feedback or revealing personal information), the experiment implements a repeated design with three rounds.

In the first part of the analysis, I focus on agents' decisions made before any feedback or public information is revealed. Overall, both men and women substantially overestimate their relative standing. This overplacement—roughly equivalent to moving from the top half to the top third of the distribution—persists across treatments and when controlling for absolute rank and academic characteristics. In the *Control* condition, where agents anticipate neither private feedback nor public observability of belief accuracy, women rank themselves about 1.3 positions lower than men on the 18-point scale. Relative to the overall male–female average, this difference equals roughly half of total overplacement.[6] This baseline pattern adds to earlier evidence that women are more moderate in their self-evaluations than men (e.g. Exley and Kessler, 2022). Building on this baseline, introducing the *Public* condition—where belief accuracy becomes observable to principals alone—markedly reduces stated rank beliefs. Both men and women lower their assessed placement by about 1.6 positions, roughly a 70% reduction in overplacement. Despite this symmetric treatment effect, the baseline gender gap in self-assessments remains largely intact, which increases the incidence of underplacement among women. In contrast, the anticipation of private accuracy feedback alone (*Private*) leads to smaller adjustments, suggesting a limited role for internal self-image concerns. Taken together, these results show that social-image concerns influence how individuals present themselves, even in settings without competitive pay schemes.

In the second part of the analysis, I focus on agents' decisions made after information about their

[4] BLESS is the Bologna Laboratory for Experiments in Social Science at the University of Bologna.

[5] The design minimised competitive payoffs to abstract from pecuniary rivalry among participants. Behavioural responses can therefore be attributed, in isolation, to exposure to belief accuracy—either privately (feedback to the agent) or publicly (scrutiny by the principal).

[6] Overplacement is defined as the difference between an agent's perceived and actual rank; lower numeric ranks indicate better relative performance. In the *Control* condition, agents on average overplaced their rank by almost three positions, so the 1.3-position gender gap corresponds to roughly 45% of this baseline overconfidence.

self-assessment accuracy in a previous round was privately or publicly revealed. After receiving private feedback (*Private*), both women and men strongly adjust their stated rank beliefs to a similar extent.[7] However, this overall similarity conceals gender-asymmetric responses to different types of feedback. Women respond significantly more strongly than men to feedback indicating overestimation—on average correcting 102% of the previous error versus 61% among men (p=0.035). By contrast, men respond more strongly to feedback indicating underestimation—83% of the previous error versus 33% among women (p=0.015).[8] Yet, the experience of public observability alone (*Public*) produces no significant changes across rounds, suggesting only minor change in social-image concerns over time. Exploratory analyses suggest that combining public exposure with private feedback—the *Joint* condition—mitigates the gender asymmetries in feedback responsiveness, with men becoming more responsive to signals of prior overestimation and women to underestimation.[9] In sum, these findings suggest that men and women are equally responsive to feedback on accuracy, yet differ in how they interpret signals of having previously over- or underestimated themselves—differences that may shape confidence formation and persistence in competitive settings.

Together, the findings offer new insights into how social aspects of competition carry psychological costs that shape competitive behaviour. Conceptually, this reframes competitive settings as belief-exposure environments that entail self- or social-evaluative concerns, reflecting context-driven responses rather than fixed predispositions (Bénabou and Tirole, 2006; Kőszegi, 2014). Empirically, the results shed light on the mechanisms behind potential gender differences in competitive behaviour by disentangling feedback and observability effects, illustrating how they may be driven by self- and social-image concerns and evolve over time. The findings contribute to the literature on gender and competition by clarifying how information exposure shapes confidence and self-evaluation—beyond material stakes. The first set of results shows a large, gender-symmetric reduction in overplacement from publicly exposing self-accuracy, which contrasts with previous studies examining observability of different types of competitive information: Buser et al. (2021b) find limited behavioural responses, Ludwig et al. (2017)[10] report effects mainly among women, and Brandts et al. (2020) document effects primarily among men when ranks are exposed to male peers. The second set of results shows direction-specific asymmetries in feedback incorporation, with women adjusting more after overestimation and men after underestimation. These patterns add to evidence that gender shapes reactions to outcome feedback (Buser and Yuan, 2019; Möbius et al., 2022) and underscore the importance of belief accuracy itself. Even in low-stakes laboratory conditions, these effects are substantial enough to move many women into underplacement, suggesting that real-world environments with stronger reputational ties or greater network visibility may amplify them. More broadly, the results indicate that confidence and self-evaluations observed in competitive environments can be understood through exposure and

[7]When aggregating feedback types indicating over-, accurate-, and underestimation of previously stated rank beliefs, no clear gender differences emerge in overall responsiveness. On average, both groups adjust their subsequent rank assessments by roughly four-fifths of the previous error, in its direction.

[8]These results remain when accounting for round-to-round changes in task performance (p = 0.038), indicating that the updating of stated beliefs stems from processing accuracy feedback rather than from relative performance itself.

[9]In the *Joint* condition, gender differences in responses to both over- and underestimation feedback narrow substantially and become statistically insignificant, compared with the *Private* condition.

[10]The experimental setup in Ludwig et al. (2017) differs from this study, primarily because participants in their design could directly assess their own accuracy *ex post*. This feature is most comparable to the *Private* and *Joint* conditions in my experiment.

feedback mechanisms, abstracted from strategic considerations or pecuniary risk preferences.

In professional settings, my findings may speak to multiple stages of job assignment through the role of social exposure and feedback in how individuals self-promote and represent themselves. If generalised, such behaviours have broad implications for career advancement and the efficient allocation of talent within and across organisations. The results demonstrate that public scrutiny can adjust self-confidence downward—sometimes to the point of underplacement—while accuracy feedback guides how individuals recalibrate their self-assessments. In professional life, accurate self-evaluation is a valuable skill: employees who misjudge their performance risk being over- or under-assigned and are costlier to manage. Both forms of miscalibration carry consequences—overly cautious self-assessments can limit applications and promotions, while excessive ones can erode credibility and distort organisational trust. The findings show that self-beliefs are dynamic and socially contingent, so even subtle variations in observability or feedback can shape how individuals evaluate themselves over time. However, salience and visibility vary in real situations: applying for jobs, negotiating pay, or presenting work to peers reveal information to different audiences, each associated with distinct reputational stakes and network ties. Experimental estimates likely capture a lower bound of these effects. Designing evaluation and feedback systems with attention to these dynamics can help reduce misallocation and unintended disparities in advancement. Future work could test these mechanisms in real networks, examining how exposure and feedback operate when evaluations are visible to colleagues, supervisors, or potential recruiters.

The remainder of the paper is organised as follows. Section 2 reviews the related literature. Section 3 presents the conceptual framework, showing how information environments affect agents' rank-assessments over time. Section 4 describes the experimental design, and Section 5 the empirical approach. Section 6 presents and discusses the findings and Section 7 concludes with limitations and policy implications.

2 Related Literature

2.1 Theoretical Literature Strand

While classical economic models often focus on monetary incentives as the primary driver of behaviour, a growing literature has highlighted the importance of non-monetary motivators such as status, esteem, and social recognition (Frey, 2007; Ball et al., 2001; Akerlof and Kranton, 2000). As Ellingsen and Johannesson (2007) argue, individuals value being respected and esteemed by others, and such social preferences can play a powerful role in shaping behaviour even in the absence of financial incentives. More broadly, status itself can be a source of utility—shaping effort, behaviour, and self-perception—even when it carries no material consequences (Ball and Eckel, 1998; Heffetz and Frank, 2011). This had led to richer models of behaviour in competitive environments that account for psychological and social costs.

Competitions as social information devices. Competitive settings do more than allocate rewards—they also produce information. Participants form beliefs about their relative performance, which can then be confirmed, contradicted, or made visible to others. This process makes competitions inherently informational and introduces the potential for psychological costs, particularly when feedback reveals miscalibrated self-views or social expectations. This mechanism has been central to theoretical models of image concerns. Bénabou and Tirole (2006, 2011) show that individuals derive utility not just from outcomes, but also from beliefs about themselves and how they are perceived. Bodner and Prelec (2003) propose the concept of "diagnostic utility," where individuals value coherence in their self-narrative and experience disutility from disconfirming signals. Grossman and van der Weele (2017) formalise the idea of strategic ignorance: individuals may avoid diagnostic feedback to protect self-image. This aligns with dual-self models such as Sautmann (2013), in which a rational self allows emotionally motivated distortions to persist.[11] Relatedly, Mijović-Prelec and Prelec (2010) argue that self-deception can serve as self-signalling: individuals manage beliefs not just to influence others, but to uphold internal identity. Schwardmann et al. (2022) extend this logic by showing that public belief reporting can feed back into private belief formation, as individuals adjust internal confidence through mechanisms of self-persuasion. These frameworks emphasise the internal costs of dissonance (Festinger, 1957) and the role of belief management in preserving self-concept.

My paper builds on this literature by shifting focus: rather than centering on outcomes or status, it examines the costs of having the accuracy of one's beliefs about relative standing being exposed. These costs may arise not only from being seen as wrong, but also from internal discomfort when identity-reinforcing beliefs are contradicted (Bénabou and Tirole, 2016). This introduces a distinct form of self- and social-image risk embedded in competitive environments—especially when beliefs are reported and subsequently revealed to be misaligned with performance.

Recent work has deepened our understanding of belief dynamics in the face of ego-relevant feedback. Zimmermann (2020) develops a dynamic model of motivated beliefs in which individuals internalise good news more than bad, leading to persistent overconfidence. This theoretical asymmetry has inspired a growing body of empirical work on belief updating and feedback processing, discussed further below. Ewers and Zimmermann (2015) further demonstrate that image concerns can lead to strategic misreporting when the truth would reflect poorly on the individual. Relatedly, Dana et al. (2007) find that individuals exploit informational ambiguity to maintain a positive self-concept—a form of moral self-protection often described as exploiting "moral wiggle room." These findings support a broader view in which belief accuracy is managed not just for performance, but to protect internal and external identity.[12]

Differentiating ego threat from status fear. Another dimension of social exposure involves status-based discomfort—particularly in contexts where individuals are publicly ranked relative to others (Tran and Zeckhauser, 2012; Ball and Eckel, 1998; Weiss and Fershtman, 1998). Psychological theory also supports the idea that epistemic exposure—the experience of having one's misbeliefs

[11]Some models of ego utility, such as Sautmann (2013), interpret self-deception or belief manipulation as the product of a multi-dimensional self-concept (e.g., emotional self vs. rational self). In that view, the separation between "self-image" and "social-image" may be more conceptual than structural. The current framework treats these costs as separable, partially to align with experimental variation (private vs. public feedback), though they may ultimately be interpreted as different facets of the self.

[12]This perspective aligns with broader theories of self-concept and ego regulation, including work on self-image as socially constructed (Rosenberg, 1965), the affective cost of dissonance and inconsistency (Baumeister, 1999; Festinger, 1957), and the motivational role of fixed versus malleable self-theories (Morton and Dweck, 2003).

revealed in front of others—carries distinct emotional and cognitive costs. Duval and Wicklund (1973) argue that objective self-awareness triggers discomfort when self-perceptions are contradicted by public information. Tracy and Robins (2004) emphasise that self-conscious emotions such as shame and embarrassment are especially likely to arise when perceived flaws are publicly exposed. Dunning (2011) adds that individuals are often unaware of their own incompetence (the Dunning-Kruger effect), making them particularly vulnerable to disconfirming feedback. These ego-related risks are amplified when others are watching. Public feedback can make overestimation socially costly, leading individuals to avoid exposure or disengage from tasks that require belief expression. Moreover, Else-Quest et al. (2012) highlight gender differences in self-conscious emotions (*SCEs*), with women reporting higher levels of shame and guilt—suggesting that the psychological cost of belief exposure may be systematically higher for women. Together, this literature supports the idea that public misjudgment imposes a distinct psychological burden—especially in competitive settings where individuals are required to "bet" on their own rank and face the social or internal cost of being proven wrong.[13]

Surveillance costs and feedback loops. From an institutional perspective, the psychological costs of belief exposure have important implications. In many organisational settings, managers rely on employees' self-assessments to allocate tasks or anticipate performance. When beliefs are systematically biased (for example, due to overconfidence or underconfidence), these judgments can lead to misallocation: overconfident individuals may overpromise and underdeliver, while underconfident ones may avoid high-impact tasks. Monitoring belief accuracy directly is costly. It requires tracking both *ex ante* beliefs and *ex post* outcomes across time and tasks (Stigler, 1962; Prendergast, 1999). In this context, exposing belief error—by making self-assessments visible to principals or peers—may serve as a low-cost substitute for formal monitoring. However, such exposure may itself be distortionary. As Kőszegi (2014) argues, behavioural contract theory shows that non-standard preferences, such as self-image concerns, can complicate incentive design and principal inference. In models with belief-based utility, agents may distort effort choices to protect self-perceptions (Santos-Pinto, 2008), while overconfidence can heighten moral hazard risks even under optimal contracts (De la Rosa, 2011). Empirical evidence from managerial settings further suggests that belief distortion can persist over time through selective recall, limiting the effectiveness of feedback and increasing agency costs (Huffman et al., 2022). These dynamics are especially relevant in repeated settings, where individuals receive feedback over time. Asymmetric updating—particularly under social scrutiny—can reinforce or erode confidence across rounds, shaping long-run engagement and withdrawal in high-exposure environments.

2.2 The Empirical Literature Strand

A large empirical literature has documented systematic gender differences in competitive performance and decision-making. Motivated by the puzzle that gender gaps in labour market outcomes persist despite a rapidly narrowing gap in qualifications, this work has explored whether men and women behave differently in competitive environments, and whether such differences might help explain observed

[13]The experimental design and accompanying model introduced in this paper formalise this distinction by embedding belief reporting and feedback into a repeated competitive task, with varying visibility. This allows us to separately identify the anticipatory cost of possible exposure and the realised cost of being miscalibrated.

disparities (see also Niederle, 2016, for an overview). Much of the early work asked whether men and women differ in how they perform under competition—for example, whether men are more likely to thrive in high-pressure settings.[14] Several studies show that men outperform women in competitive environments, even when their performance is similar under non-competitive conditions (e.g., Gneezy et al., 2003; Gneezy and Rustichini, 2004; Gupta et al., 2005). Yet, performance differences alone cannot explain why women are less likely to enter competitions–even when performance is equal—prompting a shift in focus toward preferences for competition itself.

Material stakes and the choice to compete. The seminal experiment by Niederle and Vesterlund (2007) demonstrated this puzzle clearly: women were significantly less likely than men to choose a tournament payment scheme, even when their performance was on par under piece-rate incentives. This finding has been widely replicated and is often interpreted as evidence of a gender gap in "competitiveness"—defined as a lower willingness among women to select into competitive environments that remains when controlling for performance, risk preferences, and overconfidence.[15] A number of studies suggest that the observed gender gap in willingness to compete is highly context-dependent. For instance, Dreber et al. (2011) find that the gap narrows substantially in same-gender groups, while Healy and Pate (2011) show that introducing team-based competition increases women's entry and reduces the overall gap. Dargnies (2012) further demonstrate that competition structure affects both genders: women are more likely to enter when matched with teammates, while high-performing men sometimes opt out due to concerns over partner performance. Similarly, Kuhn and Villeval (2015) and Cassar and Rigdon (2021a,b) find that women prefer team-based or pro-socially framed competitions. Other work shows that modifying incentive structures or removing risk can eliminate the gap entirely (e.g., Wozniak et al., 2016; Buser et al., 2021a). These findings problematise the idea that women are intrinsically less competitive and suggest that observed differences reflect context-sensitive responses to risk, framing, and perceived exposure.[16]

While many experimental studies rely on pecuniary stakes and winner-take-all tournament structures, it is perhaps not surprising that willingness to compete is shaped by risk preferences and confidence. In a careful decomposition, the Van Veldhuizen (2022) shows that once risk preferences and overconfidence are explicitly analysed, the observed gender gap in tournament entry largely disappears. This finding suggests that what appears to be a gender gap in "competitiveness"—as a separable trait—may instead reflect underlying individual differences in how people assess risk or perceive their own ability. This critique does not undermine earlier findings, but it reframes them: the choice to compete may say less about intrinsic competitiveness, and more about how people weigh personal exposure under pressure. One interpretation is that it is the structure of the competitive context itself that evokes these risk

[14] One early explanation focused on gender differences in task performance, particularly in *stereotypical* male tasks such as mathematics and abstract reasoning (e.g., Benbow and Stanley, 1983; Hyde, 2005; Spencer et al., 1999).

[15] The seminal work by Niederle and Vesterlund (2007) has been extensively replicated. While many studies affirm the robustness of gender differences in competitive choices (e.g., Reuben et al., 2017; Apicella et al., 2017; Halko and Sääksvuori, 2017; Reuben et al., 2017; Dohmen and Falk, 2011; Sutter and Glätzle-Rützler, 2015; Almås et al., 2016), critiques such as Gillen et al. (2019) point to the role of measurement error and argue that differences in risk preferences and overconfidence may fully account for the observed disparities.

[16] While many classic experiments include multiple stages–such as practice rounds or incentive switches—they typically focus on a single selection decision or treat competitiveness as a static trait. Fewer studies track belief dynamics over time or examine how repeated feedback and public observability interact to shape updating.

assessments and self-judgments—aligning behavioural responses with the demands of the incentive environment.

Competitive environments involve both incentives and comparison: they reward performance, but also produce information, not only about how individuals rank relative to others, but also about who chooses to compete, and under what conditions. While economic research has predominantly focused on incentives, recent work has begun to examine how the exposure inherent in comparative settings shapes behaviour and belief formation. Because of the comparative feature, competitions generate feedback about both participation and relative standing, which can carry social or psychological costs—whether the feedback is exposed to others or only observed by the individual, and even when it is not directly tied to monetary rewards.

The social risk aspects of competitions. A number of recent studies have begun to attempt to isolate the effects of social exposure—through public observability or private feedback—on behaviour in competitive environments, focusing on the role of visibility and feedback apart from material incentives. In a setting without "rivalry for resources," Schram et al. (2018) find that score rankings made visible to a peer reduce women's performance but not men's, suggesting that gender differences in sensitivity to social evaluation may affect behaviour. With a similar non-competitive reward structure, Brandts et al. (2020) find that when the choice to be ranked is imposed and the rank observer is male, men are more likely than women to prefer status ranking. However, no significant gender differences are observed when the rank observer is female, or when the choice to be ranked is private and optional. Ludwig et al. (2017) study a setting without competitive payoffs, in which participants' self-assessments are observable to a third party, who also learns whether the participant outperformed a randomly assigned peer. They find that women lower their self-assessments under observability, while men do not, suggesting that even limited exposure to evaluative feedback can affect belief reporting differently by gender. In a related design, Haeckl (2022) finds that when self-assessments are made public but performance feedback remains private, women increase their stated beliefs, while men's responses remain unchanged. This pattern suggests that observability may lead to self-enhancement, especially when the risk of external judgment is low. In contrast, Buser et al. (2021b) find that public observability of tournament decisions and outcomes does not significantly alter the gender gap in the willingness to compete. Their null result is well-identified and robust across several treatments, suggesting that social-image concerns may not influence entry choices in all competitive settings.

A note on self-promotion, image concerns, and self-stereotyping. These gendered responses to social exposure may reflect deeper differences in how individuals manage self-image and anticipate social evaluation. Several studies highlight systematic gender gaps in self-promotion and belief reporting, even outside of competitive settings. Exley and Kessler (2022) show that women are less likely to promote their own performance, even when doing so is incentivised. In follow-up work, Exley and Kessler (2023) find that while both genders avoid ego-threatening information, men are more likely to do so when social image is at stake. Relatedly, Coffman (2014) and Egerod et al. (2022) show that social norms and internalised expectations shape how men and women communicate confidence and ability—whether in group settings or in professional applications. These patterns offer insight into why observability and social evaluation may carry asymmetric psychological costs, particularly for women.

Feedback, competition outcomes, and belief adaptation. A growing literature examines how individuals respond to feedback—particularly in competitive environments—and how these responses differ by gender. Several studies show that women are more likely than men to reduce effort or withdraw after negative outcomes. Complementing these behavioural findings, a range of studies examine how feedback influences belief formation and sustained engagement. Coffman et al. (2024) and Kogelnik (2022) show that ego-relevant feedback interacts with stereotypes to produce asymmetric belief updating: women are less likely to continue in competitive tasks even when their past performance is comparable to men's. Shastry et al. (2020) provide evidence that men and women interpret noisy feedback differently: men tend to attribute negative signals to bad luck, while women are more likely to internalise them as reflecting lower ability. Enke et al. (2023) find that men and women update beliefs differently in response to performance signals, leading to gender gaps in aggregate confidence even when ability is held constant. Berlin and Dargnies (2016) demonstrate that women's willingness to compete is primarily influenced by feedback on their own performance, whereas men are more responsive to their beliefs about the competition they will face. Alan and Ertac (2019) show that providing girls with early and informative feedback significantly increases their willingness to compete, suggesting that belief calibration can play a central role in reducing gender gaps. Similarly, Shastry and Shurchkov (2024) find that female assistant professors perceive a lower likelihood of publishing after receiving a rejection compared to their male counterparts, highlighting gendered differences in reactions to negative feedback in academic publishing. These gendered feedback responses also align with broader findings in the motivated beliefs literature, where individuals are generally more responsive to affirming than to ego-threatening feedback (e.g., Eil and Rao, 2011; Coutts, 2019).

Repeated interactions. A related strand of work explicitly examines how these feedback responses unfold across rounds and over time. Buser (2016) and Buser and Yuan (2019) find that women are less likely to re-enter tournaments after underperformance, while Gill and Prowse (2014) show that losses in repeated competitions have persistent negative effects on women's performance. Kang et al. (2024) extend these findings to a high-stakes educational context, showing that women are less likely than men to retake competitive entrance exams following failure. These findings suggest that feedback—especially when tied to self-image or social evaluation—can have lasting effects on confidence and task engagement. The present study contributes by embedding belief reporting into a repeated task structure, distinguishing between the anticipation and experience of feedback, and minimising monetary stakes to isolate the psychological costs of miscalibration and exposure.

Beliefs, competitiveness, and labour market sorting. The behavioural patterns observed in lab settings may have meaningful implications for labour market outcomes. For instance, Cortés et al. (2023) find that gender gaps in job search behaviour and wage expectations emerge early in the career and are linked to differences in beliefs about ability and success probabilities. Buser et al. (2024) and Buser et al. (2014) show that laboratory measures of willingness to compete predict both long-run labour outcomes and early career-track choices, particularly among men. Flory et al. (2015)

provide field experimental evidence that framing job advertisements as competitive significantly reduces application rates among women. From a theoretical perspective, Santos-Pinto (2012) shows that overconfident individuals may be more effective at signalling ability in the labour market, contributing to wage compression and potentially explaining observed gender pay gaps. Together, these studies suggest that gender differences in belief formation, competitive preferences, and feedback sensitivity can compound over time, contributing to persistent disparities in job sorting, earnings, and advancement, even when ability is equal.

The following section sketches a stylised model in which agents form beliefs about their relative performance and may incur ego or social-image costs when those beliefs are revealed to be inaccurate. The model introduces heterogeneity in feedback sensitivity and belief updating across individuals, allowing for a flexible but tractable representation of motivated belief formation under social evaluation.

3 Conceptual Framework and Stylised Model

This section presents a stylised conceptual model intended to contextualise and interpret key features and results from the experimental setup. Non-standard modelling choices are discussed and motivated as they arise, highlighting their relevance to the experimental context. The framework focuses on agents who experience two types of psychological costs: self-image concerns, arising from realising one's own belief-miscalibrations, and social-image concerns, stemming from appearing over- or underconfident to others. The setup exhibits a largely passive principal: she may observe the agent's self-assessment and the realised outcome, but does not set material incentives, wage contracts, or monitor effort. In this respect, it departs from standard principal-agent frameworks, and ensures that the central tension lies in belief management rather than in moral hazard. Instead, the structure adapts core ideas from behavioural contract theory—extending classical models to incorporate belief-based utility linked to self-assessment accuracy and social visibility.[17]

The model proceeds in three steps. We begin with a one-shot environment (in Section 3.3) to show how mismatch costs influence an agent's self-reported rank. We then incorporate heterogeneity in sensitivity to information structures and in prior beliefs (in Section 3.5), allowing for differences across individuals or groups. Finally, we consider a repeated setting with feedback and exposure across rounds (in Section 3.6), in which agents may update beliefs only partially—reflecting motivated reasoning and selective assimilation of negative signals (e.g., Sautmann, 2013; Bénabou and Tirole, 2002). This structure is meant to clarify the psychological costs associated with being proven wrong, and to motivate the design of the experiment described later.

3.1 Basic Setting and Concepts

Environment and Output We consider a discrete-time environment with two types of players: a passive principal and a set of N decision-making agents. Each agent $i \in \{1, \ldots, N\}$ produces an

[17] See Kőszegi (2014) for a survey of behavioural contract theory, where agents derive utility from expectations and reference points over material outcomes. In contrast, the present model focuses on belief-based utility related to self-assessment accuracy and social visibility, without assuming contractible effort or explicit wage-setting. While both approaches emphasise psychological frictions, the current framework models belief exposure as an endogenous feature of competitive environments rather than as part of a contract menu.

output according to $y_i = f(a_i) + \theta_i$, where $a_i \in \mathbb{R}$ denotes the agent's intrinsic ability, and θ_i is a stochastic noise term independently drawn from a common distribution $F(\cdot)$. The production function $f : \mathbb{R} \to \mathbb{R}$ is continuous with bounded derivatives and is strictly increasing in its argument. The noise term θ_i is independent of a_i and satisfies $\mathbb{E}[\theta_i] = 0$. Agents are assumed to exert maximal effort, normalised to one, so that the variation in output arises solely from differences in ability and random shocks.[18] Agents know the functional form $f(\cdot)$ and the distribution $F(\cdot)$ governing noise, but possess only subjective beliefs about their own ability a_i. The principal remains passive and does not set incentives or influence production. Critically, the principal does not observe individual outputs y_i, but may observe agents' reported beliefs and their realised relative ranks after outcomes are determined. Agents derive utility not only from monetary payoffs linked to output, but also experience psychological costs associated with self-image and social-image concerns, which will be formalised below.

Uncertainty and Relative Ranking. Each agent observes only her own output y_i, neither the outputs nor the abilities of the other $N-1$ agents. Based on the outputs of all agents, each agent is assigned a performance rank $R_i \in \{1, 2, \ldots, N\}$, where $R_i = 1$ denotes the best rank (lower numerical value), corresponding to the highest output. Agents face uncertainty about their relative rank due both to incomplete information (on the production of others) and to random noise in production. For tractability, ranks are treated as discrete ordinal variables, and ties occur with probability zero given the continuity of the output distribution.

Belief Formation: Priors and Posterior Updating. Before observing any performance signal, each agent i holds a subjective prior belief over her own possible realised rank, denoted by r^* for $r \in \{1, \ldots, N\}$. These priors may be biased, reflecting the agent's self-perceptions rather than the objective distribution of abilities or outcomes across agents. After observing her own realised output y_i, the agent updates her belief about her relative performance. Formally, if $\mathcal{L}(y_i \mid r)$ denotes the likelihood of observing output y_i conditional on having rank r, then the posterior belief $\pi_i(r)$ is given by Bayes' rule:

$$\pi_i(r) = \frac{\mathcal{L}(y_i \mid r)\pi_i^0(r)}{\sum_{k=1}^{N} \mathcal{L}(y_i \mid k)\pi_i^0(k)}.$$

Here, the denominator sums over all possible ranks $k \in \{1, \ldots, N\}$, ensuring that the posterior $\pi_i(r)$ forms a proper probability distribution over ranks. However, posterior beliefs may be biased, either because the agent started from a subjective prior, as defined above, or because the updating process is not fully rational in a Bayesian sense; for example, agents may selectively assimilate evidence by overweighting flattering signals or discounting negative ones (Sautmann, 2013). In any case, $\pi_i(r)$ captures the agent's final subjective belief about her realized rank R_i.

Belief Reporting and Earnings. After forming posterior beliefs $\pi_i(r)$ over possible ranks $r \in \{1, \ldots, N\}$, each agent i is required to report a subjective belief distribution $\{I_i(r)\}_{r=1}^{N}$. The reported distribution must satisfy the standard properties: it must be normalised, $\sum_{r=1}^{N} I_i(r) = 1$, and assign non-negative weights $I_i(r) \geq 0$ for all r. Material earnings are determined by a proportional confidence-based

[18] Effort is assumed to be maximal and normalised across agents. This assumption abstracts from strategic effort provision and reflects the model's focus on belief formation rather than material incentive problems.

payoff structure. Specifically, agent i's monetary payoff is given by:

$$\text{Earnings} = y_i \times I_i(r^*),$$

where r^* denotes the agent's realised true rank based on actual outputs. Thus, agents earn more the higher the confidence weight they place on the correct rank. It is assumed that agents are risk-neutral over monetary payoffs, and that earnings are linear in both the output level y_i and the reported confidence $I_i(r^*)$.[19] This payoff structure is chosen specifically to capture settings in which an individual's reward depends proportionally on her degree of confidence or certainty in her actual relative performance, reflecting real-life situations where self-confidence directly influences effort allocation and task engagement or rewards *via* performance-contingent bonuses.[20]

In the absence of psychological or social concerns, a rational agent who seeks to maximise expected monetary earnings would report her true posterior beliefs, setting: $I_i(r) = \pi_i(r) \quad \text{for all } r$. However, as developed in subsequent sections, exposure to private or public feedback may generate psychological costs associated with being proven wrong, potentially distorting reporting behaviour away from the baseline truth-telling strategy. The proportional confidence-based earnings structure employed here departs from standard proper scoring rules typically used in belief elicitation experiments. A detailed discussion of alternative payoff structures and the rationale for the current design is provided in Appendix A.2. For simplicity, it is assumed—and equivalently imposed in the experimental design—that the reported belief distribution $\{I_i(r)\}_{r=1}^N$ admits a unique modal rank, denoted by $\widehat{r}_i = \arg\max_r I_i(r)$, which will play an important role in later sections when modelling psychological exposure costs.

Modal Belief Guess and Exposure. While an agent may internally assign comparable subjective probabilities to several ranks, the experimental design explicitly requires agents to report a belief distribution in which a single rank dominates, denoted by $\widehat{r}_i$. Formally, the modal belief is defined as

$$\widehat{r}_i = \arg\max_r I_i(r) \quad \text{with} \quad I_i(\widehat{r}_i) > I_i(r) \quad \text{for all} \quad r \neq \widehat{r}_i.$$

This reported modal guess serves as a clear and psychologically salient summary of the agent's subjective self-assessment: subsequent private feedback or public exposure may confirm or contradict it, potentially exposing errors in self-perception. The focus on a single modal belief, rather than the full subjective distribution, reflects a deliberate experimental design choice. Requiring agents to state a "best guess" mirrors real-world settings where individuals are often asked or pressured to make clear, publicly observable predictions or assessments despite underlying uncertainty.[21]

[19]While the experimental design allows participants flexibility in allocating confidence across ranks—potentially reflecting both subjective uncertainty and risk attitudes—the theoretical model assumes risk-neutral agents who maximise expected monetary payoffs directly. This simplification isolates the psychological costs of belief exposure without conflating them with classical monetary risk preferences.

[20]This proportional payoff structure differs from standard accuracy-based incentives or scoring rules typically used in belief-elicitation tasks. The choice here reflects the experimental focus on self-confidence and psychological exposure rather than solely accuracy or precision.

[21]Focusing explicitly on the modal belief guess as the observable belief is non-standard relative to models that consider full belief distributions or the expected value of beliefs. This choice is motivated by experimental settings and real-world environments where individuals are expected to produce single best guesses, thereby making this belief psychologically and socially salient.

Information Regimes. Before choosing $I_i(r)$, the agent knows whether she will receive private feedback about her true rank r_i^* (incurring a self-image cost if $r_i^* \neq \widehat{r}_i$), and whether the principal will observe both $\widehat{r}_i$ and r_i^* (incurring a social-image cost). The principal remains passive: she does not reward or punish performance but may observe mismatches. In anticipation of feedback or exposure, the agent may shift her stated belief weight toward ranks perceived as more modest or safer, to reduce the risk of self- or social embarrassment.

Timeline of the Game. The interaction unfolds across several stages:

> *Stage 0: Information Structure Known.* The agent learns whether she will receive private feedback about her true rank r_i^* and whether the principal will observe both her reported belief $\widehat{r}$ and realised rank r_i^*.
>
> *Stage 1: Output Realised.* The agent observes the realisation of her output y_i and forms posterior beliefs about her relative rank, but does not observe her true realised rank r_i^*.
>
> *Stage 2: Belief Reporting.* The agent reports her belief distribution $\{I_i(r)\}_{r=1}^N$, including the modal guess $\widehat{r}_i = \arg\max_r I_i(r)$, possibly factoring in anticipated exposure risks.
>
> *Stage 3: Private Feedback* (*Self-exposure*). If applicable, the agent receives private feedback about her realised rank r_i^*, which may generate self-image costs if $r_i^* \neq \widehat{r}_i$.
>
> *Stage 4: Public Observability* (*Social-exposure*). If applicable, the principal observes both the agent's reported belief $\widehat{r}_i$ and realised rank r_i^*, potentially triggering social-image costs.

Anticipation and Exposure. The agent is forward-looking. When choosing how to report her belief distribution, she weighs the benefit of concentrating belief mass on likely ranks (to maximise payoff) against the risk of exposure—privately or publicly—if her modal guess $\widehat{r}_i$ is incorrect. This trade-off underlies the central mechanism of the model. In the repeated version, discussed later, the agent also updates based on prior mismatch outcomes, allowing experience to shape belief distortion over time.

3.2 Payoff Components and Overall Utility

Monetary Payoff. The agent allocates belief confidence across N possible ranks, denoted $\{I_i(r)\}_{r=1}^N$. If the true rank is r_i^*, she receives a monetary payoff of

$$U^m(r_i^*) = y_i \cdot I_i(r_i^*),$$

where y_i is her observable output. We assume risk neutrality and linearity: each additional unit of belief weight placed on the true rank increases the payoff proportionally. This means that even if the agent's modal guess $\widehat{r}_i = \arg\max_r I_i(r)$ is incorrect, she can still earn a payoff as long as $I_i(r_i^*) > 0$.

Self-Image (Ego) Cost. If the agent receives private feedback about her true realised rank r_i^*, she may experience a self-image cost when her stated belief $\widehat{r}_i$ is incorrect. The mismatch cost takes the form:

$$U^{\text{ego}}(r_i^*) = -\Big[\alpha^+(r_i^* - \widehat{r}_i)^2 \cdot \mathbb{1}\{r_i^* > \widehat{r}_i\} + \alpha^-(\widehat{r}_i - r_i^*)^2 \cdot \mathbb{1}\{r_i^* < \widehat{r}_i\}\Big] \cdot \mathbb{1}[\text{Feedback}_A], \tag{1}$$

where α^+ governs the cost of overestimation and α^- the cost of underestimation, and $\mathbb{1}[\text{Feedback}_A]$ indicates whether private feedback occurs. Setting $\alpha^+ = \alpha^- = \alpha$ implies a symmetric version where mismatched is perceived as equally costly regardless of its direction.

Social-Image Cost. If the agent's belief report $\widehat{r}_i$ and true rank r_i^* are publicly revealed to the principal, she incurs a social-image cost when they differ. This cost arises even if the observer has no power, as the disutility stems from being seen as miscalibrated. As with ego costs, we allow for directional asymmetry in social-image penalties:

$$U^{\text{social}}(r_i^*) = -\Big[\gamma^+(r_i^* - \widehat{r}_i)^2 \cdot \mathbb{1}\{r_i^* > \widehat{r}_i\} + \gamma^-(\widehat{r}_i - r_i^*)^2 \cdot \mathbb{1}\{r_i^* < \widehat{r}_i\}\Big] \cdot \mathbb{1}[\text{Observed } (\widehat{r}_i, r_i^*)],$$

where γ^+ and γ^- represent sensitivity to being seen as overconfident or underconfident, respectively.

Overall Utility. The agent's *ex post* utility given true rank r_i^* is:

$$U^W(r_i^*) = U^m(r_i^*) + U^{\text{ego}}(r_i^*) + U^{\text{social}}(r_i^*),$$

where each term may be specified in symmetric or asymmetric form.[22]

Ex-Ante Expected Utility. After observing y_i but before learning r_i^*, the agent decides on a belief distribution $\{I_i(r)\}_{r=1}^N$ to report (and thus $\widehat{r}_i = \arg\max_r I_i(r)$) to maximise expected utility. Taking expectations over r_i^* conditional on y_i, distributed according to $\pi_i(r)$:

$$\mathbb{E}\Big[U^W(r_i^*) \,|\, y_i\Big] = \sum_{r=1}^{N} \pi_i(r)\, U^W(r).$$

This reflects the trade-off between monetary gain (*via* $I_i(r_i^*)$) and the anticipated psychological or social mismatch costs, conditional on the feedback regime.

3.3 A Single-Period Model: Anticipation and Optimal Self-Assessment

We begin with a single-period environment in which the agent observes her output y_i and forms subjective beliefs $\pi_i(r)$ over possible ranks $r \in \{1, \ldots, N\}$. Based on these beliefs, she chooses a belief distribution $\{I_i(r)\}_{r=1}^N$ to report, satisfying $\sum_r I_i(r) = 1$, and a modal guess $\widehat{r}_i = \arg\max_r I_i(r)$. Her payoff depends on how well her reported beliefs match the true (but unknown) rank r_i^*, and whether any mismatch is revealed to herself or to others. All components and parameters follow Section 3.2.

The Agent's Problem. After observing y_i and forming subjective beliefs $\pi_i(r)$ over possible ranks $r \in \{1, \ldots, N\}$, the agent decides on a belief distribution $\{I_i(r)\}_{r=1}^N$ to report (and thus chooses a

[22]For notational simplicity, we suppress explicit dependence of U^{ego} and U^{social} on $\widehat{r}_i$.

modal guess $\widehat{r}_i = \arg\max_r I_i(r)$) to maximise expected utility. Her objective is:

$$\max_{\{I_i(r)\}} \quad \mathbb{E}_{r_i^* \sim \pi_i} \left[U^W(r_i^*) \mid y_i \right] \quad \text{s.t.} \quad \sum_r I_i(r) = 1, \quad I_i(r) \geq 0 \quad \forall\, r \in \{1, \dots, N\}, \quad \widehat{r}_i = \arg\max_r I_i(r)$$

where $U^W(r_i^*) = U^m(r_i^*) + U^{\text{ego}}(r_i^*) + U^{\text{social}}(r_i^*)$ and its components are defined as follows:

$$U^m(r_i^*) = y_i \cdot I_i(r_i^*)$$

$$U^{\text{ego}}(r_i^*) = -\mathbb{1}[\text{Feedback}_A] \cdot \begin{cases} \alpha^+(r_i^* - \widehat{r}_i)^2 & \text{if } r_i^* > \widehat{r}_i \\ \alpha^-(\widehat{r}_i - r_i^*)^2 & \text{if } r_i^* < \widehat{r}_i \\ 0 & \text{if } r_i^* = \widehat{r}_i \end{cases}$$

$$U^{\text{social}}(r_i^*) = -\mathbb{1}[\text{Observed } (\widehat{r}_i, r_i^*)] \cdot \begin{cases} \gamma^+(r_i^* - \widehat{r}_i)^2 & \text{if } r_i^* > \widehat{r}_i \\ \gamma^-(\widehat{r}_i - r_i^*)^2 & \text{if } r_i^* < \widehat{r}_i \\ 0 & \text{if } r_i^* = \widehat{r}_i \end{cases}$$

Interpretation. The agent balances three considerations:

1. *Expected payoff:* She aims to maximise expected monetary payoff by reporting higher confidence on ranks she believes are more likely to be correct.
2. *Self-image cost:* If feedback is private and $\widehat{r}_i \neq r_i^*$, she anticipates that she will experience discomfort if she turns out to be wrong about herself.
3. *Social-image cost:* If the modal guess is publicly observed and mismatched, she anticipates that she will face embarrassment or reputational damage.

The trade-off is shaped by the asymmetry parameters α^+, α^- (ego) and γ^+, γ^- (social). Overplacement ($r_i^* > \widehat{r}_i$) may be more costly than underplacement ($r_i^* < \widehat{r}_i$), or vice versa.

Characterising the Optimal Guess. Faced with a trade-off between monetary payoff and exposure risk, the agent evaluates each possible modal guess $\widehat{r}_i$ and selects a belief distribution $I_i(r)$ that maximises her expected utility. Since the rank space is discrete, this results in a finite-choice optimisation problem.

$$\begin{aligned} \mathbb{E}_{r_i^* \sim \pi_i} \left[U^W(r_i^*) \mid y_i \right] = \sum_{r=1}^{N} \pi_i(r) \cdot \Big[& y_i \cdot I_i(r) \\ & - \mathbb{1}[\text{Feedback}_A] \cdot \left(\alpha^+(r^* - \widehat{r})^2 \cdot \mathbb{1}\{r^* > \widehat{r}\} + \alpha^-(\widehat{r} - r^*)^2 \cdot \mathbb{1}\{r^* < \widehat{r}\} \right) \\ & - \mathbb{1}[\text{Observed } \widehat{r}] \cdot \left(\gamma^+(r^* - \widehat{r})^2 \cdot \mathbb{1}\{r^* > \widehat{r}\} + \gamma^-(\widehat{r} - r^*)^2 \cdot \mathbb{1}\{r^* < \widehat{r}\} \right) \Big] \end{aligned}$$

Even without a closed-form solution, some comparative statics are immediate:

- If α^+ or γ^+ increases, the agent fears *overplacement* more and chooses a higher (worse) $\widehat{r}_i$.
- If α^- or γ^- increases, she fears *underplacement* more and chooses a lower (better) $\widehat{r}_i$.
- If feedback is private only, ego costs dominate; if public, social-image concerns may cause strategic

underreporting.

Formally, the agent's utility function is continuous and defined on a compact simplex of belief distributions. With a consistent tie-breaking rule for modal beliefs, existence of an optimal solution follows by standard arguments (see Appendix A in Section A.3 for details).

3.4 Comparative Statics: The Effect of Mismatch Sensitivities

We summarise the directional effects of increasing mismatch sensitivity parameters on the agent's optimal modal guess $\widehat{r}_i^*$.

Proposition 3.1 (Ego Cost Sensitivity). *Suppose feedback is private. Then:*

- *Increasing α^+ (the cost of overestimation) weakly increases $\widehat{r}_i^*$: the agent becomes more conservative, avoiding high (overconfident) guesses.*
- *Increasing α^- (the cost of underestimation) weakly decreases $\widehat{r}_i^*$: the agent becomes more assertive, avoiding low (underconfident) guesses.*

Proposition 3.2 (Social-Image Sensitivity). *Suppose the agent's guess $\widehat{r}_i$ is publicly observed. Then:*

- *Increasing γ^+ (the cost of appearing overconfident) weakly increases $\widehat{r}^*$.*
- *Increasing γ^- (the cost of appearing underconfident) weakly decreases $\widehat{r}_i^*$.*

These comparative statics reflect the core trade-off in the model. When mismatch costs rise in one direction (e.g., overplacement), the agent shifts her modal guess away from that region, preferring to err in the less costly direction. This framework thus predicts directional shifts in self-assessment as mismatch sensitivities change.

Discussion. This one-shot model shows how the agent's stated belief $\widehat{r}_i$ reflects a trade-off between maximising monetary payoff and avoiding reputational or psychological penalties. When mismatch costs are symmetric, the agent's guess may align closely with the mode of her belief distribution $\pi_i(r)$. When the costs are asymmetric, she shades her guess toward the safer direction—lowering $\widehat{r}$ if underplacement is cheap, or raising it if overplacement is especially costly. In what follows, I introduce heterogeneity in mismatch sensitivity. For instance, if compared to men, women are more concerned about being seen as overconfident (i.e., higher γ^+), or respond more negatively to internal feedback (i.e., higher α^+), then this framework predicts more conservative self-assessments. Such asymmetries may explain observed gender differences in willingness to compete, exposure aversion, and belief updating.

3.5 Gender Heterogeneity and Biased Priors

Thus far, the model has assumed that all agents share the same mismatch sensitivities and prior beliefs. We now introduce gender-based heterogeneity along the two dimensions: sensitivity to private and public information costs, as well as prior beliefs about rank.[23]

[23] Formal versions of these results, including full propositions and proof sketches, are provided in Appendix A.5.

Mismatch Sensitivities. Suppose men and women differ in their psychological response to exposure. For example, if men are more averse to privately discovering they have overestimated their ability (i.e., higher α_m^+), they may select safer guesses to avoid internal disappointment. Conversely, if women are more sensitive to appearing overconfident in public (i.e., higher γ_f^+), they may shift their self-assessment downward when observability is expected. These asymmetries shift the optimal modal belief $\widehat{r}_i$ differently under public and private feedback regimes.

Proposition 3.3 (Gendered Sensitivity to Feedback)**.** *Suppose men and women have identical beliefs $\pi(r)$ but differ in exposure sensitivities. If $\alpha_m^+ > \alpha_f^+$ or $\gamma_f^+ > \gamma_m^+$, then men are more conservative when private feedback is expected, while women are more conservative when exposure is public. That is, $\widehat{r}_m^* > \widehat{r}_f^*$ in the former case and $\widehat{r}_f^* > \widehat{r}_m^*$ in the latter.*

Biased Priors. Alternatively, gender differences in self-assessment may stem from distinct belief distributions. Suppose men place more mass on top ranks (i.e., optimistic priors), while women have more conservative expectations. Even if mismatch costs are identical, posterior beliefs $\pi_m(r)$ and $\pi_f(r)$ will differ after observing performance. This implies systematic differences in belief allocations and modal guesses.

Proposition 3.4 (Gender Differences from Prior Beliefs)**.** *Suppose $\pi_m^0(r)$ stochastically dominates $\pi_f^0(r)$ (in the first-order sense). Then, even under identical mismatch costs, men will select a better (lower) $\widehat{r}^*$ on average than women.*

Summary. Together, these two sources of heterogeneity—belief sensitivities and prior confidence—can explain why men and women may report systematically different self-assessments in competitive environments, even when performance is similar. These predictions can be tested directly by comparing modal beliefs across feedback regimes and between genders.

3.6 A Repeated Exposure Model and Belief Updating

We now extend the model to a repeated environment with $T = 3$ rounds, aligning with the experimental design. Each round follows a fixed five-stage structure, allowing belief formation, exposure, and partial updating across rounds. The core trade-off—between monetary gain and mismatch exposure—remains present, but now unfolds over time:

> **Stage 1: Prior Belief for Round t.** At the beginning of round t, the agent holds a belief distribution $\pi_t(r)$ over possible ranks. In round $t = 1$, this belief reflects only her initial expectations, possibly biased. In later rounds ($t > 1$), this prior incorporates past performance and any mismatch feedback from previous rounds.
>
> **Stage 2: Production and Observation.** The agent produces output y_t, observes her own performance, but not that of others. Her true rank r_t^* remains unknown at this stage.
>
> **Stage 3: Belief Update and Guess.** After observing y_t, the agent updates her belief to $\pi_t'(r)$ and selects a belief distribution $I_t(r)$ over ranks. Additionally, she reports a modal guess $\widehat{r}_t = \arg\max_r I_t(r)$, anticipating that it may later be exposed to herself or to an external party. Mismatch sensitivities α and γ are stable across rounds.

Stage 4: Feedback Realisation. Depending on the feedback regime, the agent may observe her true rank r_t^* (private feedback), and/or a third party may observe the pair $(\widehat{r}_t, r_t^*)$ (public feedback). If mismatch is revealed, the agent incurs a private or social cost, as described in Section 3.2.

Stage 5: Partial Update for Next Round. If $t < T$, the agent adjusts her prior for round $t+1$ based on:

– The performance signal y_t

– The mismatch $r_t^* - \widehat{r}_t$ (if feedback was received), using partial update weights

Special Case: Round 1. In the first round, the agent has not yet received any mismatch feedback. Her belief $\pi_1(r)$ reflects only her initial (possibly biased) prior, shaped by self-perception and expectations. From round 2 onward, beliefs are influenced by observed mismatch and performance changes from earlier rounds. In this conceptual setup, each round t effectively mirrors the one-shot problem, but now the initial beliefs π_t reflect prior mismatch signals and performance changes. A full backward-induction approach could solve for the agent's equilibrium choices in each round; however, we focus on this partial-adjustment process to highlight how mismatch and new performance signals shape self-assessment over time in a transparent manner. Following the partial-update rule:

$$\widehat{r}_{t+1} = \widehat{r}_t + \eta^+ \cdot (r_t^* - \widehat{r}_t) \cdot \mathbb{1}\{r*_t > \widehat{r}_t\} + \eta^- \cdot (r_t^* - \widehat{r}_t) \cdot \mathbb{1}\{r_t^* < \widehat{r}_t\} \tag{2}$$

where $\eta^+, \eta^- \in [0,1]$ are asymmetry parameters for overplacement and underplacement feedback. These weights may differ across agents or groups.

Implications. This repeated structure shows how mismatch exposure alters belief paths:

- *Consistent overestimation* (i.e., $r_t^* > \widehat{r}_t$) leads to upward drift in $\widehat{r}_t$ if $\eta^+ > 0$
- *Consistent underestimation* leads to lower $\widehat{r}_t$ if $\eta^- > 0$
- *Group differences* in η^+ or η^- produce divergent belief paths across rounds
- *No feedback* implies no belief correction; the agent repeats her prior belief trajectory

Formal lemmas and propositions supporting these comparative dynamics are provided in Appendix B.

Signal Weighting. In addition to mismatch feedback, the agent may also use changes in y_t to guide beliefs. Let $\delta^y \in [0,1]$ capture the relative weight placed on own performance vs. feedback. This allows for agents who trust their own output more than social comparisons, or *vice versa* (see Section B.3 in Appendix B).

Discussion. This repeated-round extension shows how belief paths evolve under asymmetric learning. Even if mismatch costs remain fixed, agents with low η^+ or high δ^y may remain miscalibrated over time. The framework provides a clean basis for comparing behaviour across feedback regimes and across groups.

3.7 Key Insights and Testable Predictions

3.7.1 One-Shot Model Predictions

The one-shot model yields four testable implications about how agents report their beliefs about relative standing (including selecting $\widehat{r}_i$) under different exposure regimes and psychological cost structures. These predictions follow from the comparative statics in Section 3.4 and the gender heterogeneity in Section 3.5. While the model allows for psychological costs in both directions—penalising agents for overplacement (α^+, γ^+) and underplacement (α^-, γ^-)—the predictions below focus on the overplacement margin (Prediction 1 and 2).[24] This reflects both the empirical prevalence of overplacement in the data, and the structure of decision-making in the model: agents report beliefs *ex ante*, and can shift their report only in one direction. Thus, only one mismatch type can be behaviourally relevant at the margin for a given agent. The model also permits gender heterogeneity in mismatch sensitivities, allowing cost parameters such as α^+ and γ^+ to differ between men and women. The predictions below derive implications from these differences, rather than assuming them (Predictions 3 and 4).[25]

- **Prediction 1 (Effect of social exposure).** When belief accuracy is expected to be socially salient (i.e., impose a social-image mismatch cost γ^+), agents report higher (i.e., more conservative) ranks to reduce the anticipated reputational cost of appearing overconfident.
- **Prediction 2 (Effect of self-exposure).** When belief accuracy is expected to be privately revealed to the agent (i.e., impose an ego-image mismatch cost α^+), agents report higher ranks to reduce the anticipated discomfort of learning they were overconfident.
- **Prediction 3 (Gendered sensitivity to social exposure).** If women are more sensitive than men to observable miscalibrations (i.e., $\gamma_f^+ > \gamma_m^+$), then under conditions of social exposure, women will report higher (i.e., more conservative) ranks than men, for equivalent posterior beliefs.
- **Prediction 4 (Gendered sensitivity to self-exposure).** If one gender is more sensitive to internal miscalibration (e.g., $\alpha_f^+ > \alpha_m^+$), then under private feedback, that group will report more conservative (i.e., higher) ranks than the other.
- **Prediction A (Belief gaps from priors).** Differences in reported beliefs may also arise from group-level differences in prior expectations (e.g., $\pi_m^0(r) < \pi_f^0(r)$). If one group holds more optimistic priors about their rank, it will report lower (better) modal rank beliefs, even if mismatch sensitivities are identical.

The predictions above highlight the model's core mechanism: belief reporting is shaped not only by perceived likelihoods, but also by the psychological costs of being revealed as miscalibrated—whether privately or publicly. We now turn to a multi-round version of the model, where feedback and exposure accumulate across periods, potentially reinforcing or mitigating these effects over time.

[24] Focusing on α^+ and γ^+ allows for a clean mapping between theoretical mechanisms and the observable belief distortions in the experimental data.

[25] That is, they specify what follows if one group is more sensitive than the other to a given type of mismatch.

3.7.2 Repeated-Round Model Predictions

The repeated-round extension generates four predictions about how belief reports evolve over time in response to feedback exposure. These predictions arise from the structure of partial belief updating and its interaction with realised performance and feedback about prior belief accuracy.

- **Prediction 5 (Directional learning from repeated feedback).** If agents partially incorporate feedback about prior mismatch (i.e., $\eta^+ > 0$ or $\eta^- > 0$), then repeated exposure to directional signals will shift beliefs over time. In particular, agents who consistently overestimate their rank ($r_t^* > \widehat{r}_t$) will revise their self-assessments upward, while those who repeatedly underestimate ($r_t^* < \widehat{r}_t$) will revise downward.
- **Prediction 6 (Asymmetric updating by feedback types).** If agents weigh overplacement and underplacement signals differently (i.e., $\eta^+ \neq \eta^-$), belief revisions will be asymmetric. For example, if $\eta^+ > \eta^-$, agents will revise more strongly in response to signals of overplacement than underplacement.
- **Prediction 7 (Gendered responsiveness to feedback types).** If updating weights differ across groups, then belief revisions will vary systematically by gender and feedback type. For example, if incorporation of overplacement feedback is stronger for women than men ($\eta_f^+ > \eta_m^+$), and of underplacement feedback stronger for men than women ($\eta_m^- > \eta_f^-$), belief adjustments will reflect these asymmetries.
- **Prediction 8 (Feedback visibility moderates belief updating).** If agents anticipate that feedback will be publicly observable, social-image concerns may reduce the extent of belief updating—particularly for those with high sensitivity to overplacement in public ($\gamma^+ > 0$). As a result, even when informative feedback is received, public exposure may dampen adjustments that would otherwise occur under private conditions.

These predictions help distinguish between behavioural types and feedback regimes. They also provide dynamic implications that can be directly tested in the experiment by tracking changes in $\widehat{r}_t$ across rounds.

Summary of Theoretical Framework. The model highlights how agents manage their belief reports when exposed to feedback about relative performance. In the one-shot case, they trade off expected payoff against potential mismatch costs—internal (ego) and external (social). When feedback is repeated over rounds, belief paths evolve through partial updating, and the direction and magnitude of adjustment depend on both exposure and asymmetry in learning. Heterogeneity in mismatch sensitivity and prior beliefs produces systematic variation across individuals or groups. The resulting predictions provide a transparent link between the structure of feedback, the agent's internal calculus, and observable behaviour—offering a foundation for empirical analysis in competitive settings.

4 Experimental design and data generation

This section details a novel experimental design and its implementation, with particular focus on the timeline of tasks, the belief elicitation mechanism, and the structured variation in information conditions. The experimental setup is closely aligned with the key mechanisms formalised in the conceptual model: agents form and report beliefs about their ranks, including a "best guess" (representing their strongest rank belief), and face information treatment conditions in which the accuracy of their self-assessment is *ex post* either revealed to themselves privately, made publicly visible to an external observer, both, or withheld entirely. The design isolates the psychological costs of belief exposure, allowing a careful examination of how agents adjust their self-assessments under different conditions and how these evolve dynamically over repeated experimental rounds. Importantly, the design is free from competitive payment schemes, ensuring that decisions reflect belief management motives rather than strategic responses to monetary incentives.

While the experimental design closely mirrors the informational structure formalised in the conceptual model, it necessarily abstracts from several behavioural and contextual complexities. First, the model assumes that agents are risk neutral with respect to monetary outcomes, which simplifies analysis but does not account for potential heterogeneity in risk preferences—something the experiment can partially capture, for example through participants' behaviour in the incentivised lottery task. Second, the model treats performance as exogenous and abstracts from effort choices, whereas participants in the experiment complete real-effort tasks that may elicit variable motivation and engagement. Third, the model presumes full internalisation and understanding of the information conditions, whereas in practice, participants may misunderstand or interpret public and private information exposure in noisier or more heterogeneous ways. These simplifications are deliberate, allowing the model to isolate core psychological mechanisms, while the experiment tests their empirical relevance in a more behaviourally rich environment.

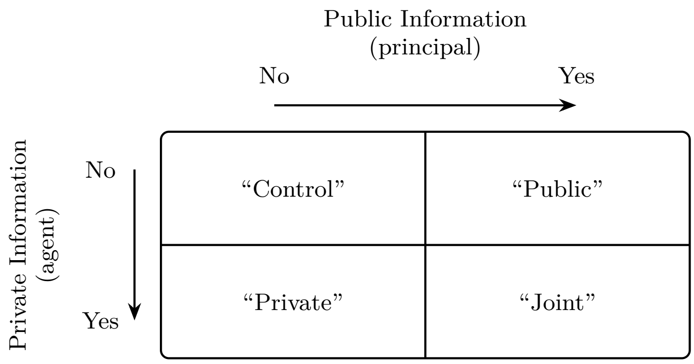
The remainder of this section outlines the experimental setting, treatment conditions, and procedures. Section 4.1 describes the experimental structure, including participant roles, groupings, and the information conditions. Section 4.2 summarises administrative procedures and data collection, and Section 4.3 briefly reflects on methodological considerations.

4.1 Experimental Setup and Treatments

This subsection sets up the experimental environment and describes its structure, including participant roles and the information conditions that define the four treatment arms used in the design. In the experimental environment, there are two participant roles. Each session consists of 24 participants, with 18 randomly assigned to serve as *agents*, and 6 to serve as *principals*. Three agents are randomly matched with one principal to form a group, without any recruitment or selection process. Roles and groups remain fixed throughout the experiment. Principals play a structurally important but passive role, acting as non-strategic observers of agents' decisions and behaviours. This paper focuses exclusively on the behaviour of agents.

The experiment adopts a between-subject design, with agents randomly assigned to one of four treatment conditions. These conditions vary whether and to whom information is provided that allows the accuracy of an agent's rank belief to be verified against their true performance rank. When this information is revealed to the agent, it constitutes private feedback; when it is revealed to the agent's principal, it becomes public information. The treatment conditions are designed to test how self-image and social-image concerns correspond to two distinct dimensions of information dissemination, as outlined in the stylised model. Specifically, the treatment conditions vary both the availability of information and its recipient: whether—and to whom—information is disclosed that enables the agent's self-assessment accuracy to be verified against their realised performance rank. When this information is revealed to the agent herself, it constitutes private feedback; when revealed to the agent's principal, its public nature makes it constitute social information. This results in a fully crossed 2×2 design with four distinct treatment arms, as illustrated in Figure 1 below.

Figure 1: Treatment matrix: Information on agent self-assessment accuracy by role and condition



NOTES: "Yes" and "No" indicate whether an agent's rank self-assessment accuracy is observable, and by whom. Treatment conditions: "Control" = {Private Feedback = 0; Social Information = 0}, "Private" = {1, 0}, "Public" = {0, 1}, "Joint" = {1, 1}. All treatment arms are equal in size.

The experiment is organised as a sequence of three identically repeated rounds, each comprising four structured stages: (1) a real-effort task in which agents solve decoding problems; (2) elicitation of rank beliefs; and, after the introduction of treatment variations, (3) agent-specific feedback; and (4) submission of information to the agent's principal. All instructions were presented using a labour market framing, delivered individually on screen, and participants advanced through the stages one at a time.[26] Each stage was introduced separately, and no part of the experiment could be completed without first reading the corresponding instructions in full.[27] A complete set of translated instructions are available in Appendix J.

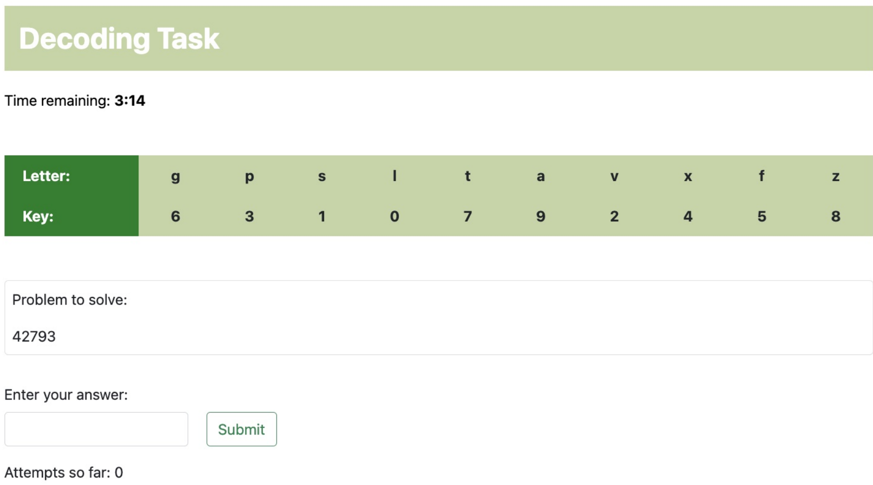
Stage 1: The real-effort task. To establish agent performance, each round begins with a real-effort *decoding* task in which agents have four minutes to solve as many problems as possible. Each problem consists of translating (i.e., decoding) a 5-digit number into a corresponding 5-letter string. An example is shown in Figure 2. Agents earn one point toward their total score for each correct solution; incorrect answers carry no penalty and cannot be changed once entered. The task score also serves as the basis

[26] Participant roles (detailed below) were framed in labour market terms: agents were referred to as "Employees", and principals retained their title. Agents selected among "Contracts" to state their rank beliefs, determining a performance-based "wage", and so on.

[27] Each part of the experiment was preceded by a short unpaid practice task (2 and 4 minutes, respectively), followed by a comprehension check covering the key rules and mechanics of Stages 1 and 2.

for agents' earnings (explained below). At the end of the task, agents are informed of the number of problems they attempted and submit a guess of how many they solved correctly.[28] Then, the performance scores of all 18 agents in a session are ranked. Rank 1 is assigned to the agent with the highest task score, rank 2 to the agent with only one other agent performing better, and so on, until rank 18 is assigned to the agents having 17 other agents scoring better (i.e., with the lowest performance). This implies that agents with tied scores share the better (lower-numbered) rank.[29] Subsequently, agents receive private feedback on their score, but no information about their relative performance or the performance of others.

Figure 2: The Real-Effort Task—A Decoding Problem



NOTES: The figure shows an example of a decoding problem. The displayed problem "42793" decodes to the solution "xvtap" (answers are not case-sensitive). The interface displays the remaining time on-screen and increases the count of attempted problems after each submitted answer. No other feedback is provided.

Stage 2: Rank belief elicitation. The second stage elicits agents' beliefs about their possible performance-based ranks (as determined in Stage 1) and determines their experimental earnings. In each round, agents are endowed with 19 units of experimental currency (ECUs), which are non-transferable and non-storable. Agents undertake a self-assessment task by deciding how to allocate their endowment across a set of 18 possible ranks. They may allocate the entire endowment to a single rank, or spread it over multiple ranks—provided that all 19 ECUs are used. The allocation is completed in two steps:

Step 1: Agents select one or more ranks—numbered 1 to 18—to which they intend to allocate some of the endowment, subject to the constraint that they must choose at least one and at most eighteen ranks.

Step 2: Agents allocate ECUs across the ranks selected in Step 1, subject to the condition that one rank must receive at least one ECU more than any other. This ensures a single most-preferred

[28] A correct guess is incentivised and rewarded with experimental currency equivalent to 1 euro.

[29] Ties are resolved using a dense ranking rule, For example, if the two highest-scoring agents both have 17 correct answers, both receive rank 1. The next-best agent is then assigned rank 3, since *two* agents scored higher. If no further ties occur, the lowest rank assigned remains 18.

rank.[30]

The Step 2 constraint is designed to impose a unimodal distribution over agents' reported beliefs, thereby yielding an interpretable measure of the rank each agent believes to be most likely—presented to them as their "most preferred contract". Agents were informed *ex ante* of the information structure tied to their treatment condition—that is, whether they or their principal would observe the outcome of their rank belief report. This structure was common knowledge from the outset and shaped expectations, and potentially also decisions, already in the first round.

Experimental earnings and incentive structure. Agents' earnings depend directly on whether they accurately identify their realised performance rank through the allocation of ECUs. If an agent allocates ECUs to their actual rank, their earnings are calculated as the number of ECUs allocated to that rank multiplied by their task score from Stage 1. Allocating no ECUs to the true rank results in zero earnings. This structure ensures that agents are rewarded for accurate self-assessments while allowing them to express uncertainty. Allocating all ECUs to a single rank maximises expected earnings only if the agent is sufficiently confident that this rank matches their true performance. Distributing ECUs across multiple ranks reduces payoff variance and allows agents to express subjective uncertainty or accommodate risk preferences. The design permits a range of belief allocation strategies.[31] Principals' earnings are directly tied to the earnings of their agents: by default, each principal earns one third of the total earnings of each of their three agents.[32] While each round was rewarded according to the same structure, only one round was randomly selected for payment. To prevent hedging across rounds, this was announced only at the conclusion of the experiment. The final two stages—Stages 3 and 4—implement the information treatments introduced by the treatment structure (Figure 1).

Stage 3: Agents' private feedback. In the third stage, agents receive individual feedback in the form of a round summary table. It is shown at the end of each round and displayed privately on screen. In all conditions, the *baseline feedback* information includes the agent's own task score from Stage 1 and a table listing the selected ranks from Stage 2 alongside the number of ECUs allocated to each. The table is ordered by ECU amount in descending order, with the most preferred rank (i.e., the one receiving the highest allocation) highlighted at the top. In rounds 2 and 3, the most preferred rank from the previous round is also shown at the bottom of the screen. The specific content of the agents' feedback varies across treatments and is outlined below.

Control & Public conditions: In these conditions, the baseline feedback described above constitutes the full information agents receive. The summary table shown to agents in these conditions appears as follows:

[30] The agents can revise the rank selection in step 2 and repeat step 1 if desired.

[31] While the experimental instructions do not reference self- or social-image concerns, the design structure allows for behaviour consistent with such motives to emerge. See Section 4.3 for a detailed discussion of how the design relates to the model's assumptions.

[32] This is the default payment scheme. For a discussion of optional alternative payment structures available to principals, *see* Alamaa (2024).

Summary of round 2

This is the summary of last round, including how you selected the contract(s).
Your score last round was 19.

ECU per Contract "wage"	*selected* **Contract** "rank"	
11 ECUs	**7**	**"most preferred contract"**
4 ECUs	6	
3 ECU	8	
1 ECU	5	

In round 1: *your* ***most preferred contract*** *was 7.*

Private & Joint conditions: In these conditions, agents receive, in addition to the baseline feedback, further information regarding the accuracy of their rank selection. Specifically, the agent's true performance rank is displayed next to their selected rank. Two additional columns are included: "Difference", which shows the numerical gap between the true and selected rank (defined as true rank minus selected rank); and "Direction", which categorises this gap as either "Underestimation", "Accurate", or "Overestimation", depending on whether the selected rank was worse (i.e., a higher number), equal to, or better (lower number) than the true rank. The summary table shown to agents in these conditions appears as follows:

Summary of round 2

This is the summary of last round, including how you selected the contract(s).
Your score last round was 19.

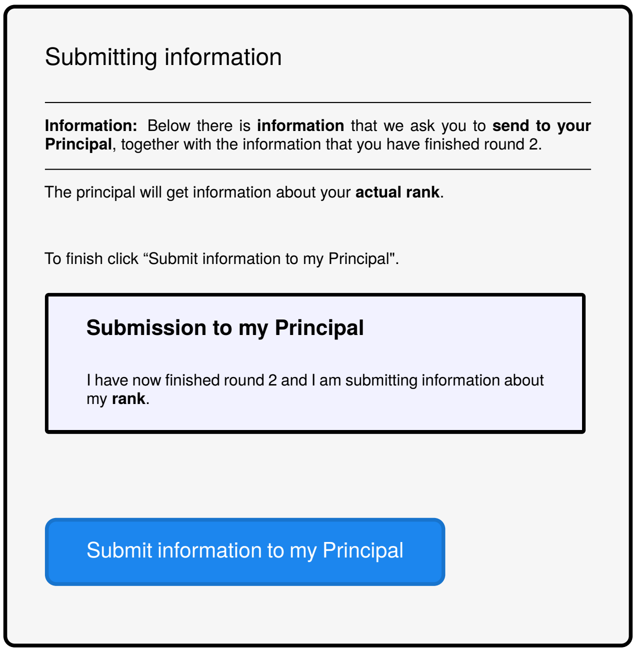
ECU per Contract "wage"	*selected* **Contract** "rank"	**Rank** (actual)	**Difference** (rank-contract)	**Direction**	
11 ECUs	**7**	**8**	**+1**	**Overest.**	**"most preferred contract"**
4 ECUs	6	8	+2	Overest.	
3 ECU	8	8	±0	Accurate	
1 ECU	5	8	+3	Overest.	

In round 1: *your* ***most preferred contract*** *was 7 and your rank was 6.*

Stage 4: Information submission to the principal. In the final stage of each round, agents submit information to their principal *via* a structured form. This form reflects treatment-specific variation on whether agents expose information on their self-assessment accuracy. In all conditions, agents disseminate *baseline information* to their principals by providing their true performance rank. This information is displayed in a principal round-summary table listing the actual ranks of all three of their employees, labelled *Employee 1–3*. In the later rounds 2 and 3, also the true ranks from previous rounds are shown below the table (regarding round 1, and 1 and 2 respectively). The additional information revealed in Stage 4 varies with the treatment condition, as described below.

Control & Private conditions: In these conditions, the true rank is the only information that agents submit. Thus, principals receive no information that would allow the agent's rank belief to be compared to their actual performance. The submission interface used in these conditions is shown below in Figure 3.

Figure 3: Submission interface shown to agents in Stage 4, by treatment condition



Control & Private condition

NOTES: In the *Control* and *Private* conditions, agents "silently" submit only their actual rank.

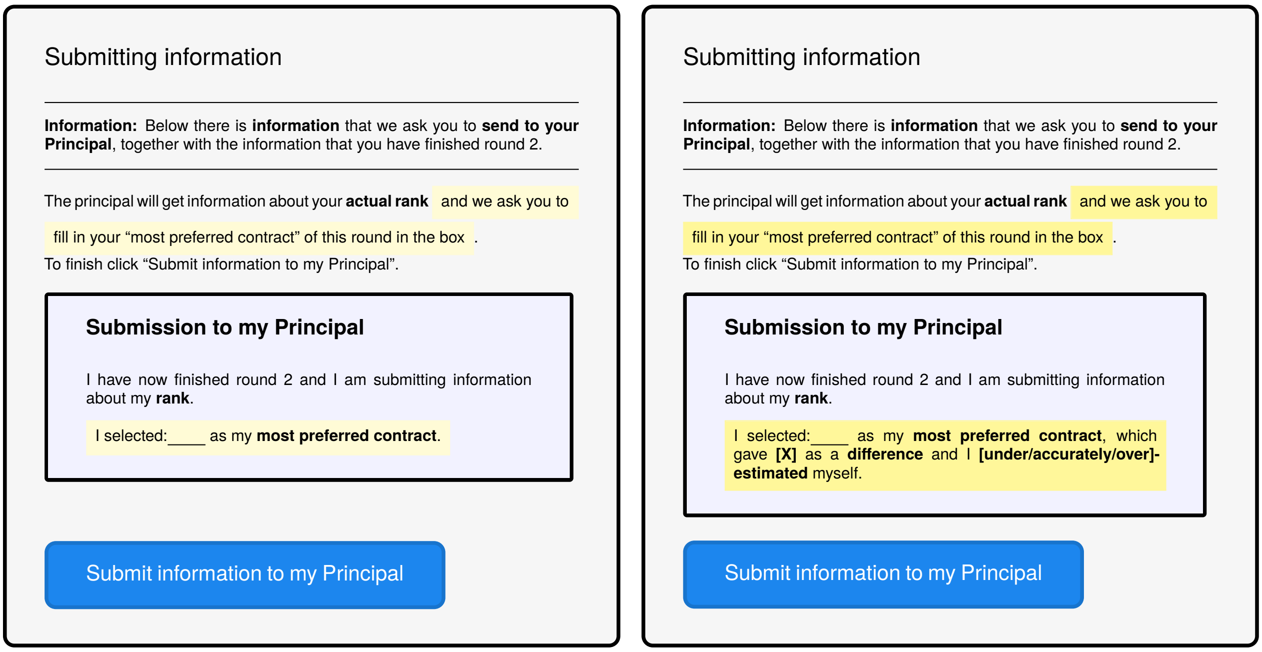
The interface is identical for principals in both conditions, and no belief-related information is visible.[33] Additional submitted information appears only in the Public and Joint conditions, as outlined below.

> *Public & Joint* conditions: In these conditions, agents submit their most preferred rank (i.e., the contract number receiving the highest ECU allocation) in addition to their true performance rank. This allows the principal to observe the agent's self-assessed rank alongside their actual performance. The principals' round-summary table displays the agent's selected rank, the true rank, and two derived indicators—"Difference", showing the numerical gap between the ranks; and "Direction", categorising the discrepancy as "Underestimation", "Overestimation", or "Accurate".

While the information received by principals is the same in the *Public* and *Joint* conditions, the salience of what is being submitted differs for the agents (*see* Stage 3). Agents in the Public condition do not know whether their rank selection is accurate, but agents in the Joint condition do. As a result, only the latter are aware of the interpretation the principal will be able to make based on their submission. The two submission form versions are shown in Figure 4.

[33] In the *Private* condition, agents are shown their true rank in Stage 3 and could, in principle, recognise what they are submitting to the principal. However, the submission interface does not explicitly display this information.

Figure 4: Submission interface shown to agents in Stage 4, by treatment condition



(a) *Public* condition **(b)** *Joint* condition

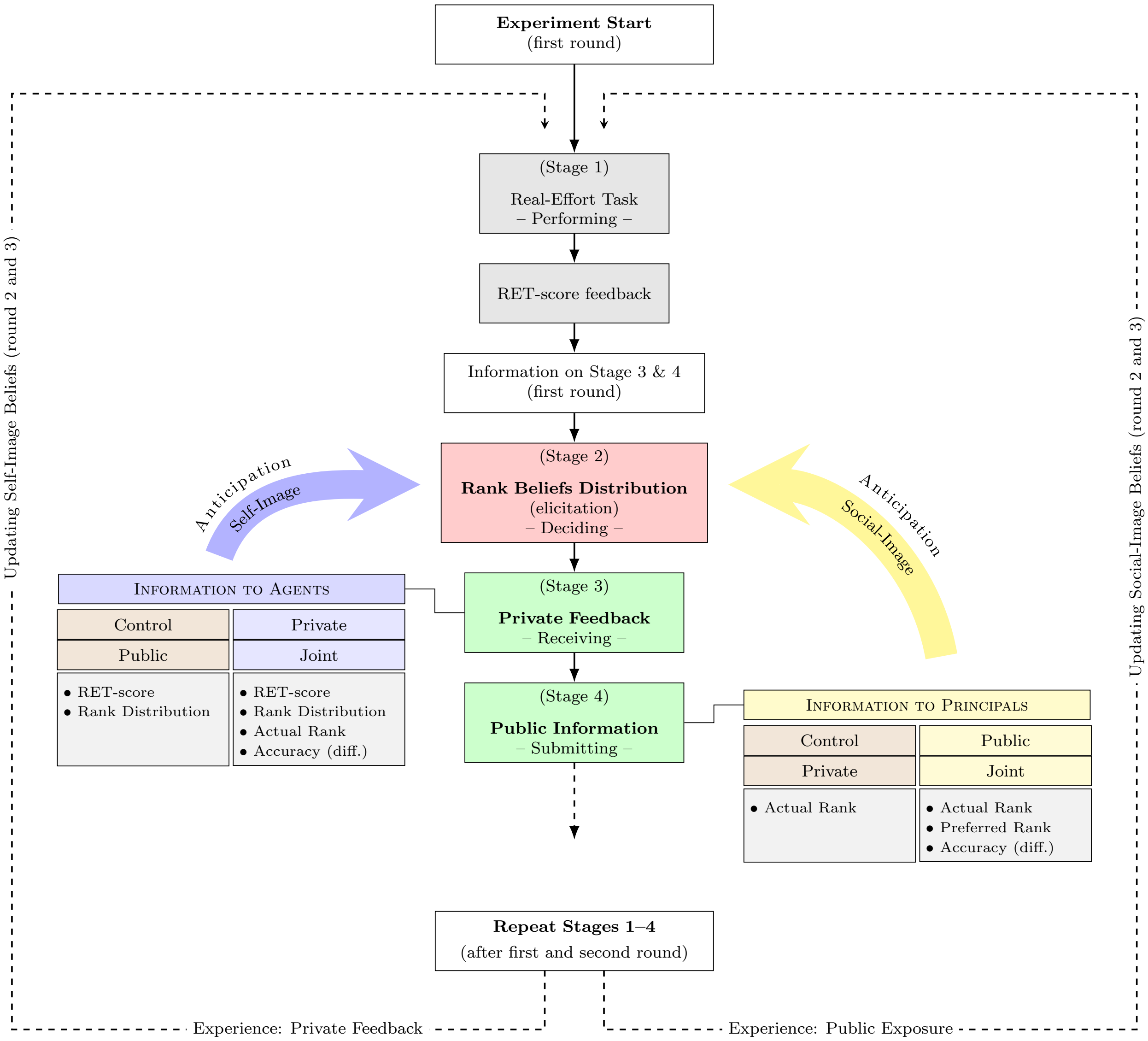
NOTES: In the *Public* condition (left), agents submit their actual rank and fill in their “most preferred contract”, corresponding to their modal rank guess. In the *Joint* condition (right), the interface additionally displays the difference and accuracy classification tied to that submission.

Repeated structure. Figure 5 illustrates the overall structure of the experiment, which is organised around an experimental round repeated three times. Each round includes the four stages described above, shown in the vertical pillar alongside the timing of RET-score feedback. The belief elicitation task is repeated within subjects, allowing rank beliefs to be collected in each round. This repeated structure enables the identification of potential differences between anticipation effects (e.g., expecting personal feedback or social exposure) and experience effects (e.g., having received such feedback in earlier rounds).

Other data-generating parts. In addition to the main experimental tasks, a brief pre-experiment survey was administered to collect demographic information, including age, gender, number of completed study years, and field of study, as well as a language inventory. The latter included a self-assessed proficiency rating for several listed languages and generated a count variable for the number of languages spoken. Study fields were later reclassified into STEM and LLL categories (outlined in Section D.3 in Appendix D). During the experimental tasks, data were also collected on the number of trial attempts in both the real-effort and belief elicitation tasks, as well as responses to a five-item comprehension check. These are primarily used for diagnostic and robustness purposes.

After the final round, participants completed an incentivised risk-preference task based on a simplified lottery choice structure inspired by Holt and Laury (2002). This was followed by a short exit survey capturing non-incentivised measures of risk attitudes and negotiation preferences, which are used in some robustness specifications or analysed in related work. The full experimental instructions and

Figure 5: Flow of the experimental stages over rounds (three)



Notes: The figure illustrates the repeated structure of the experiment, showing the four within-round stages and their sequence across three rounds. The RET score is revealed after Stage 1, and submitted information is sent to principals at the end of Stage 4. The design allows agents' experiences of personal feedback or social exposure to influence beliefs and behaviour in subsequent rounds.

decision-screens are provided in English in Appendix J.

4.2 Administration and experimental procedures

The experimental design and the main hypotheses were pre-registered at the Open Science OSF Preregistries.[34] The data generation process was aimed for a laboratory experiment. The lab experiment was conducted in the Bologna Laboratory for Experiments in Social Science (BLESS) at the University of Bologna during 2023–2024. Participants were recruited through ORSEE (Greiner, 2015), using

[34]The project is registered as Alamaa, C. (2023, June 11), "The role of the observer for self-assessment—gender differences in image-concerns" (*previous* working-title) and can be retrieved from `https://osf.io/8z39j`.

separate recruitment lists for men and women based on registered gender.[35,36] I collected data over a total of 40 experiment sessions from 768 subjects, 576 agents and 192 principals, for the four treatment conditions. Each session was planned to include 24 participants, but due to a technical error affecting one treatment group, 8 sessions were re-run.[37,38]

Details on session inclusion, randomisation, and sample restrictions are provided in Section 5.1 and 4.3. The experimental software was programmed in oTree© (Chen et al., 2016) that was run on the laboratory server. Random assignments—for example, to treatments, groups, or lottery outcomes—were handled directly in the application code. The experimental instructions were fully computerised and provided in Italian. The entire experiment took approximately one hour to complete.[39] The procedure began with reading the instructions, followed by a two-minute decoding task trial. Participants then tested the rank investment module and completed a comprehension test, which took about 25 minutes. The main experiment, consisting of three rounds, was completed in an additional 35 minutes.[40] Average earnings were 13.3 €, including a flat participation fee of 5 € or 10 €, depending on session logistics. One out of the three experimental rounds was randomly selected for payment and announced after completing the experiment. Subjects earned experimental currency units, which was converted to euros at a fixed rate of 10 ECUs = 1 €.[41]

4.3 Discussion of the experimental design

The experimental design aims to isolate the psychological effects of exposing self-assessments to either private feedback or external visibility, while ensuring simplicity for participants and producing analytically clean measures. Several design choices carry implications worth noting. First, the real-effort task was selected to balance two goals: generating sufficient within-session variation in performance, and avoiding gendered associations common in arithmetic or grammar-based tasks. I used a decoding task originally introduced by Chow (1983), which moderately achieved the first aim and successfully fulfilled the second. Second, although the primary focus is on belief accuracy, the design necessarily involves assigning agents a performance rank. This introduces a status-ranking context that may influence behaviour. Since all participants are ranked and this structure is constant across treatments, any such effects are orthogonal to the treatment variation. Prior evidence suggests limited gender differences in responses to status ranking (Brandts et al., 2020). Third, the belief elicitation module

[35] Participants were required to submit a stated informed consent, a Consent Form, prior to participating in the experiment. The Consent Form follows EU GDPR Law (2016/679) concerning storage of Personal Data.

[36] At the time of the data collection the ORSEE database of the participant pool at the University of Bologna (BLESS Lab) included about 5,700–6,300 registered subjects. These subjects were primarily university students, though not exclusively.

[37] As a result, the final dataset includes 16 sessions with 16 participants (12 agents), 16 sessions with 20 participants (15 agents), and 8 sessions with 24 participants (18 agents).

[38] To ensure full sessions, more participants were invited than needed. Typically, 32–36 individuals were scheduled per session, from which 24 were randomly selected to participate. Those not selected (recruits) received a 5 € show-up fee and were eligible to participate in a future session, but no individual participated more than once.

[39] This experiment was conducted simultaneously with another study that began *after* all decisions had been made by the participants in this experiment. The design considerations and settings for both experiments were independently configured.

[40] This time includes waiting periods due to the requirement for synchronised start times across rounds and stages.

[41] To maintain the *no-feedback* condition in the *Control* treatment, where agents could otherwise infer their performance rank from earnings, a random bonus of 0, 1, or 2 times their experimental earnings was added to their final payment (excluding the score guess worth 1 euro).

involves a relatively dense interface, requiring participants to allocate all 19 ECUs across ranked contracts under specific constraints. Although contract numbers map directly onto performance ranks, interpreting these as ordinal may not have been fully intuitive to all participants. Despite interface supports, I cannot exclude the possibility that some belief reports reflect framing effects or cognitive confusion. Fourth, the social-exposure treatments are necessarily stylised. Participants do not interact or observe one another directly; instead, informational exposure occurs via submitted data. While this differs from studies using physical or verbal exposure (e.g., Buser et al., 2021b; Schram et al., 2018), the approach allows for tight control over the information environment and was chosen in part due to the study's gender focus. Fifth, the payoff structure was designed to avoid negative externalities across roles. Agents' earnings do not reduce those of their principals and vice versa, helping to limit the influence of other-regarding preferences such as altruism or inequality aversion. Sixth, the design opens the possibility for strategic behaviour in which an agent deliberately performs poorly to reduce uncertainty in belief reporting—e.g., by aiming for last place and allocating all ECUs to the bottom-ranked "contract". While rare, some agents appear to have pursued this strategy.

5 Data and Empirical Strategy

This section outlines the dataset, empirical approach, and testable hypotheses. Section 5.1 describes the sample and presents key summary statistics. Section 5.2 introduces the main estimation strategy and outcome variables. Section 5.3 states the hypotheses, which are motivated by the conceptual model and, where applicable, aligned with the pre-registered analysis plan.

5.1 Sample, Randomisation, and Descriptive Statistics

This paper is based on experimental data collected from 40 laboratory sessions, involving 575 agent subjects (289 male, 286 female), matched 3:1 with 192 principal subjects.[42] Each session was designed to include 24 participants—18 agents and 6 principals—assigned across four treatment conditions. Groups were formed by assigning either 3 or 6 agents (i.e., 4 or 8 participants in total) to each treatment arm in alternation. The randomisation protocol was predefined and designed to achieve gender balance across both roles and treatments.[43] One session was initially run under an incorrect software configuration and was subsequently rerun with the same allocation and settings.[44,45]

The analysis uses agent-level data from three experimental rounds, yielding approximately 1500 agent-round observations in total, with about 540 per round. A number of agent-round observations are excluded due to evidence of strategic distortion in task performance. A full breakdown by treatment and gender is provided in Appendix C. Table 1 shows how the base sample of 575 agents is distributed across the four treatment conditions. Both the *Control* and *Joint* conditions include 144 agents (72

[42]The minor gender imbalance resulted from two recruitment-related issues: one agent was excluded due to a mismatch between registry and declared gender, and one session oversampled a male participant due to insufficient female availability.

[43]Treatment order was determined using Random.org and communicated to the BLESS Lab in advance of data collection.

[44]Session 16 was run with the debug mode enabled, which made correct answers visible to participants. The session was rerun as session 33.

[45]Due to a configuration error in the experiment code, all sessions related to the *Public* condition were re-run in July and September 2024. The final dataset includes only sessions that were conducted under the corrected settings.

male and 72 female). The *Private* condition includes 143 agents—72 male and 71 female—due to a mismatch between database registry and declared gender that was detected *post hoc*.[46] The *Public* condition includes 144 agents, with 73 male and 71 female participants.This imbalance resulted from limited female availability during one session.

Table 1: Final data—by gender and treatment conditions

	Control	Private	Public	Joint	Total
Male	72	72	73	72	289
Female	72	71	71	72	286
Total	144	143	144	144	575

To confirm the success of random assignment, I test for baseline balance on agents' demographic characteristics across the four treatment conditions. The results, presented in Table D.2 in Appendix D, show that covariates are generally balanced across treatments: p-values from joint F-tests and pairwise comparisons typically exceed 0.10. One exception appears in the proportion of agents majoring in Languages, Linguistics, or Literature (LLL), with a joint F-test p-value of 0.048 and a significant imbalance between the *Control* and *Private* conditions ($p = 0.007$). This imbalance is addressed in robustness checks with covariate adjustment.[47] Summary statistics by gender are presented in Table D.1 in Appendix D. Across the full sample ($N = 575$), gender differences in demographic and academic characteristics are small and statistically insignificant, with the exception that women report speaking slightly more languages than men ($p < 0.001$).

5.2 Empirical Strategy

The empirical analysis focuses on agents' reported rank beliefs and their accuracy across the three experimental rounds. The unit of analysis is the agent-round. After applying the exclusion criteria, the main analysis sample includes 554 agents in Round 1 and 538 in Round 2. Round 3 is used in robustness analysis.

I estimate average treatment effects of exposure to feedback and social observability on self-assessment outcomes, using ordinary least squares (OLS) with standard errors clustered at the session level. Treatment variation is captured through two binary indicators: $E_i = 1$ if the accuracy of the agent's self-assessed rank is subject to private feedback (ego exposure), and $S_i = 1$ if it is subject to public observability (social exposure). The *Control* condition serves as the omitted category. The main specification includes round fixed effects where applicable. The baseline estimating equation is:

$$Y_{it} = \mu + \beta_E E_i + \beta_S S_i + \beta_{ES}(E_i \times S_i) + \lambda_t + \varepsilon_{it}$$

where Y_{it} is the outcome for agent i in round t, and λ_t are round fixed effects (included in specifications pooling Rounds 2 and 3). The coefficients β_E, β_S, and β_{ES} capture the effect of private feedback,

[46] One participant was registered in ORSEE as female but self-identified as male during the session; due to the mismatch, this observation was excluded from the final dataset.

[47] The same balance tests were conducted separately for the samples used in rounds 1, 2, and 3 (shown in Table D.3). Results are consistent across rounds, with only minor variation in the size and location of the LLL imbalance.

social exposure, and their interaction (i.e., the *Joint* treatment), respectively. All main specifications include controls for agents' task performance in the corresponding round (RET score), years of higher education, and indicator variables for STEM and LLL fields of study.

To examine heterogeneity in treatment effects by gender, I estimate a second specification with interaction terms:

$$Y_{it} = \mu + \sum_k \beta_k T_k + \delta F_i + \sum_k \theta_k (T_k \times F_i) + \lambda_t + \varepsilon_{it}$$

where $T_k \in \{E_i, S_i, E_i \times S_i\}$ and F_i is a binary indicator for female. In this specification, all control variables are also interacted with gender to allow for fully flexible covariate effects across groups. The main outcome variables include the agent's reported modal rank $\widehat{r}_i$ and their placement, defined as the difference between realised rank and reported modal rank. A negative value indicates underplacement, a positive value indicates overplacement, and zero reflects accurate placement. Additional measures are discussed in Appendix E.

In addition to these primary outcomes, I examine agents' ability to identify their true rank across the full belief distribution. Specifically, I measure whether any ECUs are allocated to the realised rank (extensive margin), and what share of the endowment is placed on it (intensive margin). While not central to the main analysis, these outcomes serve as complementary measures of self-assessment precision.

In the repeated rounds, I also explore how agents adjust their beliefs following feedback. This includes measuring the extent to which belief updates incorporate information about prior mismatch (e.g., over- or underplacement), and whether such adjustments differ systematically by feedback direction or gender. To formalise this, I define the feedback signal in round t as

$$\text{signal}_t = \widehat{r}_t - r_t^*,$$

where $\widehat{r}_t$ denotes the agent's modal rank belief and r_t^* her realised rank. A negative signal therefore indicates overplacement (the agent performed worse than she believed and should revise towards a worse, higher-numbered rank), while a positive signal indicates underplacement (she performed better than she believed and should revise towards a better, lower-numbered rank). To specifically evaluate gender differences in feedback adjustments (*see* Hypothesis H11), I construct a proportional adjustment variable that captures the share of the prior-round feedback signal that is reflected in updated rank beliefs. A full definition and discussion are provided in Appendix E.1. Dynamic responses are discussed in more detail in Section 6.

5.3 Testable Hypotheses

This subsection presents the main testable hypotheses, derived from the conceptual model in Section 3 and the pre-analysis plan (PAP). Hypotheses are grouped by conceptual theme: *priors*, *anticipation of belief accuracy exposure*, and *learning from realised exposure*.

Group 1: Priors and Baseline Gender Differences

H1 ***Estimation bias:*** When guessing their score (given known attempts), women are less likely to overestimate than men.

H2 ***Belief precision:*** Women allocate their belief over more ranks and invest less per selected rank, indicating lower belief precision.

H3 ***Placement bias:*** Women report higher (worse) modal ranks $\hat{r}$ than men, consistent with more conservative self-assessments.

H4 ***Actual accuracy:*** Women are as likely as men to allocate at least some belief weight on their true rank (extensive margin).

Group 2: Anticipated Exposure Effects

H5 ***Anticipated social-image consideration.*** When agents anticipate that belief accuracy will be publicly observable ($S_i = 1$), they state more conservative (i.e., higher/worse) modal ranks to avoid appearing overconfident.

H6 ***Gender difference in social-image concerns.*** Relative to men, women report more conservative (i.e., higher/worse) modal ranks when they anticipate that belief accuracy will be publicly observable, (greater sensitivity to social-image concerns).

H7 ***Anticipated self-image consideration.*** When agents anticipate that belief accuracy will be privately revealed ($E_i = 1$), they report more conservative (i.e., higher/worse) modal rank to mitigate the risk of internal misjudgement.

H8 ***Gender difference in self-image concerns.*** Relative to men, women report different modal ranks when private feedback is anticipated ($E_i = 1$); (non-directional).

Group 3: Experience and Feedback Incorporation

H9 ***Belief updating.*** Agents adjust their modal rank in response to realised feedback: overplacement leads to higher next-round guesses; underplacement to lower.

H10 ***Asymmetry in updating.*** Agents respond more strongly to overplacement than to underplacement (denoted $\eta^+ > \eta^-$ in the model's framework, *see* Section 3.6).

H11 ***Gender difference in updating.*** Women respond more strongly to overplacement feedback than men ($\eta_f^+ > \eta_m^+$); the opposite may hold for underplacement ($\eta_f^- < \eta_m^-$).

H12 ***Self- and social-exposure interaction.*** Information about a past mismatch has a stronger influence on subsequent belief updating under private exposure ($E_i = 1$) than under additional public exposure ($S_i = 1$), where social-image concerns may suppress learning.

Appendix F provides additional details on the hypotheses, including their confirmatory status and how they map onto predictions from the theoretical model.

6 Results

This section presents two main sets of results, each beginning with descriptive statistics and associated baseline hypothesis tests, which contextualise the empirical analysis reported afterwards. These descriptive insights ground the examination of how agents allocate their endowments across ranks to express beliefs under different information conditions. The first set considers behaviour in Round 1, when agents make decisions under anticipated exposure, before any information about belief accuracy has been revealed (Section 6.1). The second set examines behaviour in Round 2, after belief accuracy has been revealed in the Private, Public, and Joint treatments, when agents make decisions having experienced the exposure introduced in Round 1 (Section 6.2).

6.1 Anticipating Exposure (Round 1)

This section first describes agents' baseline performance and belief patterns (6.1.1), followed by treatment effects on stated rank beliefs and self-assessment accuracy (6.1.2). It then examines whether these adjustments reflect strategic shading or a broader shift in the underlying rank beliefs (6.1.3), and concludes with preregistered outcomes on accuracy and earnings (6.1.4).

6.1.1 Baseline Performance and Belief Patterns (Round 1)

Table 2 presents descriptive statistics for agents' first-round task performance, score beliefs, and rank selection behaviour. The upper panel summarises results from the real-effort task (RET) and belief elicitation; the lower panel covers rank selection and self-assessment accuracy.[48]

Agents scored an average of 16.72 points on the four-minute RET, with no meaningful gender difference across or within treatments. The score-based ranks were assigned within sessions, averaging 8.61 out of 18 (where 1 is the best rank and 18 the worst). Although average ranks do not differ significantly by gender, men were nearly three times as likely as women to be ranked last (7.5% vs. 2.6%; $p = 0.008$). Before learning their RET score, agents estimated how many items they solved correctly. Estimation errors were modest, and women reported significantly more conservative score guesses than men, supporting **H1**.[49]

Agents' rank-allocation decisions are presented in the lower panel of Table 2.[50] On average, agents allocated ECUs across 5.3 ranks in Round 1, while only 0–7 percent allocated all 19 ECUs to a single rank. Figure 6 illustrates the distribution of the number of ranks agents selected by gender.

Testing **H2** related to belief precision, I find no evidence that women allocate ECUs across more ranks than men; the gender difference is small and statistically insignificant (Section G.3 in Appendix G). Agents' most preferred rank—that is, the rank receiving the largest share of their ECU allocation—was on average rank 6 out of 18. This suggests that, on average, agents believed their performance placed them in the top third of the distribution.[51] Women selected significantly higher (i.e., worse)

[48]Treatment-level breakdowns are provided in Appendix G.1, Table G.1.

[49]**H1, Estimation bias:** Knowing number of attempted problems, women are less likely to overestimate their scores compared to men.

[50]Recall that these are incentivised by conditioning payoffs on the selected rank matching the true rank, with only one rank paying per agent per round.

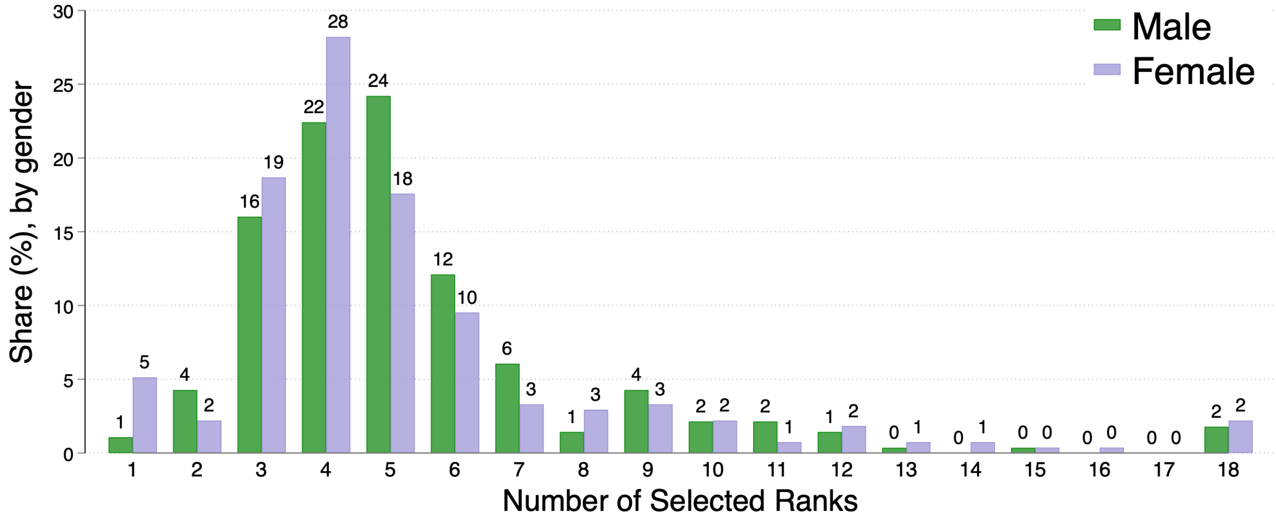
[51]In the absence of ties, the expected average rank in a uniformly distributed group of 18 is 9.5.

Table 2: Descriptive Statistics, by gender—first round

		All			
	Scale	(1) All	(2) Male	(3) Female	(4) *p*-val.
REAL-EFFORT TASK PERFORMANCE					
Score	5–32	16.722	16.911	16.527	0.197
		(3.499)	(3.926)	(2.991)	
Score belief	2–31	16.695	16.961	16.421	0.081
		(3.635)	(3.959)	(3.253)	
Estimation	-12–4	-0.027	0.050	-0.106	0.130
		(1.214)	(1.061)	(1.350)	
Underest.	0/1	0.193	0.160	0.227	0.046
Accurate est.	0/1	0.581	0.598	0.564	0.421
Overest.	0/1	0.226	0.242	0.209	0.350
Rank	1–18	8.606	8.338	8.883	0.214
		(5.153)	(5.466)	(4.804)	
Ranked first	0/1	0.074	0.082	0.066	0.474
Ranked last	0/1	0.051	0.075	0.026	0.008
RANK ALLOCATIONS					
No. of ranks	1–18	5.298	5.370	5.223	0.570
		(3.031)	(2.842)	(3.217)	
Only one	0/1	0.031	0.011	0.051	0.006
Preferred rank	1–18	6.065	5.367	6.784	0.000
		(4.076)	(3.864)	(4.168)	
Share of allocation[1]	0–1	0.363	0.345	0.382	0.008
		(0.167)	(0.139)	(0.190)	
Placement[2]	-16–16	2.542	2.972	2.099	0.078
		(5.833)	(5.658)	(5.987)	
Underplac.	0/1	0.296	0.263	0.330	0.087
Accurate plac.	0/1	0.070	0.068	0.073	0.795
Overplac.	0/1	0.634	0.669	0.597	0.079
N		554	281	273	554

[1]Agents' endowment allocations to their most preferred rank must be $\leq$ 2 ECUs, setting the lower range of approximately 0.11 (from 2/19).
[2]Placement is measured as the true rank minus the preferred rank, so that a negative (positive) number implies the underplacement (overplacement) of agent's rank and a zero signifies an accurate placement of the most preferred rank.
NOTES: This table presents descriptive statistics from the first round, pooled across all treatment conditions. The upper panel reports real-effort task (RET) performance, score beliefs, and corresponding estimation accuracy—categorised as under-, accurate-, or overestimation. The lower panel summarises rank selection behaviour, including the number of ranks selected, an indicator for selecting only one rank, the preferred (modal) rank, the share of ECUs allocated to the preferred rank, and measures of placement accuracy (difference between actual and modal rank). Columns report means separately for all agents (col. 1), men (col. 2), and women (col. 3), with *p*-values in col. 4 based on *t*-tests (for continuous variables) or χ^2-tests (for binary variables). For treatment-specific breakdowns, see Appendix G Table G.1.

Figure 6: Agents' number of selected ranks in belief elicitation, by gender—first round



NOTES: The figure shows the distribution agents' number of ranks selected to allocate endowment to, by gender. Bars are expressed as the percentage share per gender selecting each number of ranks.

preferred ranks than men (6.78 vs. 5.37; $p < 0.001$), indicating more conservative self-assessments.

To formally test Hypothesis **H3**, which posits that women underplace themselves relative to men, I regress the modal rank on gender using data from the *Control* group across all rounds. Controlling for task performance and round fixed effects, women report ranks approximately 1.5 positions lower than men on average (Table G.5, $p < 0.05$). This result is consistent with the interpretation that women are less self-confident in their relative performance, even in the absence of any information exposure or feedback.

6.1.2 Treatment Effects (Round 1)

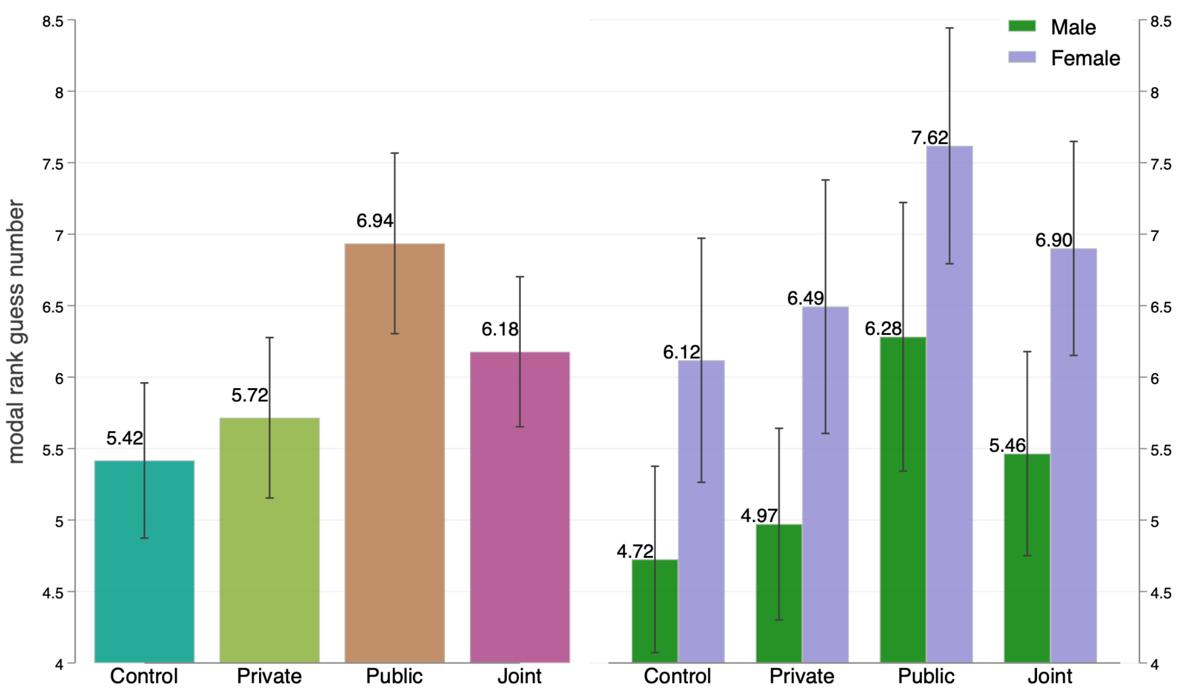
Anticipating the exposure of belief accuracy—particularly when this information will be observable to the principal—leads agents to state more conservative self-assessed ranks in the first round. Under public exposure, both men and women report lower modal rank beliefs, which results in lower overplacement relative to the *Control* condition.

Result 1: Effects on modal rank beliefs. The left panel of Figure 7 shows the average rank to which agents allocated their highest belief weight (modal guess) across treatment conditions. A higher number indicates a lower perceived standing. In all three treatments, agents report worse ranks compared to the *Control* condition. Specifically, the average rank increases by 0.3 ranks in the *Private* condition (6%, $p = 0.28$) 1.6 ranks in the *Public* condition (28%, $p < 0.01$), , and 0.8 ranks in the *Joint* condition (14% $p = 0.05$). Among these, only the shift in the *Public* condition is both statistically significant and substantial (in the raw mean comparison). This suggests that anticipated public observability reduces reported confidence: when agents know their belief accuracy will be visible to others, they adopt more conservative self-assessments. Anticipating private feedback alone (*Private* condition) has no significant effect. However, the shift observed in the *Joint* condition ($p = 0.052$)

offers tentative evidence that public visibility—even when combined with private feedback—shapes how individuals present their beliefs. This pattern is consistent with a self-presentational response to anticipated public scrutiny.

The right panel of Figure 7 breaks down the average preferred rank (modal guess) by gender across treatment conditions. On average, women consistently report worse ranks than men, indicating more conservative rank self-assessments. In the *Control* condition, women select rank 6.1 on average, compared to rank 4.7 for men—a significant difference of 1.4 ranks ($p = 0.03$). This baseline gender gap persists across all treatment arms, and there is no indication that the treatments amplify or attenuate the difference between men and women. In the *Public* condition, men exhibit a slightly larger downward shift than women in their stated rank, although this difference—based on raw means—is not statistically significant.[52] This descriptive pattern is revisited in the regression analysis below, where covariates are included.

Figure 7: Modal rank number, by treatment and by gender - first round



NOTES: The figure displays agents' mean modal rank number that they preferred (*i.e.* in the rank they allocated the most ECUs to) in the first round. To the left by treatment conditions and to the right both by treatment and gender.

Table 3 confirms these patters using OLS estimates of the modal rank as a function of treatment assignment and gender. Each column pair compares a treatment group to the Control, with the second column in each pair including an interaction between gender and treatment to test for differential effects. All specifications control for performance score and academic background variables,[53] and show that only the *Public* treatment significantly affects reported rank placement: participants assigned to

[52]These raw differences relative to *Control* are 1.56 ranks for men and 1.50 ranks for women.

[53]Including years of completed higher education and indicators for whether majoring in a STEM or LLL (Languages, Linguistics, or Literature) field.

Table 3: Modal rank beliefs, first round

	Private		Public		Joint	
	(1)	(2)	(3)	(4)	(5)	(6)
Female	1.42**	1.61*	1.24**	1.18	1.26**	1.39
	(0.51)	(0.64)	(0.40)	(0.70)	(0.39)	(0.69)
Treatment	0.26	0.22	1.56*	1.75*	0.79	0.86
	(0.48)	(0.57)	(0.63)	(0.79)	(0.49)	(0.53)
Treatment × Female		0.08		-0.14		-0.17
		(0.88)		(0.87)		(0.87)
Round score	✓	✓	✓	✓	✓	✓
Academic cont.	✓	✓	✓	✓	✓	✓
Female × Academic cont.		✓		✓		✓
Mean of dep. variable	5.57	5.57	6.2	6.2	5.81	5.81
Observations	274	274	275	275	277	277

Notes: The table shows OLS estimates of regressing the modal rank on treatment dummies; a gender dummy, and; a gender dummy interacted with the treatment indicator. Columns are paired and each pair represents one treatment condition compared to the *Control*. All models control for score performance and for academic background control variables, additionally which are interacted with the gender dummy for models with explicit gender effects, in Models (2), (4), (6). Standard errors are in parentheses, clustered at session level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

the *Public* condition state ranks that are substantially lower (i.e., worse) than in the *Control* in which agents' belief accuracy was not exposed.

Additionally, the only effect sizes of the *Public* and *Private* conditions differ significantly from one another ($p < 0.01$),[54] highlighting that public exposure—rather than private feedback—drives changes in agents' reported rank beliefs. By contrast, the differences in effect sizes between the *Public* and *Joint* conditions, as well as between the *Private* and *Joint* conditions, are statistically indistinguishable at conventional confidence levels ($p = 0.06$ and $p = 0.15$, respectively). These patterns suggest that the presence of public observability is the key mechanism influencing behaviour, rather than the combination of both exposure types.

This interpretation aligns with the conceptual framework: private feedback and social exposure might affect belief reporting through distinct psychological mechanisms. Public observability reduces overplacement, likely because agents anticipate reputational costs or feelings of social embarrassment if proven wrong. Notably, this response arises even in an environment where personal reputation is largely irrelevant—interactions are brief, anonymous, and devoid of material consequences—underscoring how even minimal visibility can meaningfully shape behaviour. In contrast, private feedback alone does not generate a comparable shift in reported rank, indicating weaker ego-related concerns in isolation.

Taken together, these findings support **Hypothesis H5** (social-image effect): agents report more conservative self-assessed ranks when anticipating public observability. In contrast, the null result for private feedback does not support **H7** (self-image effect), suggesting that anticipated ego-related

[54] Estimated effects are similar in magnitude across genders, but only the male coefficient reaches statistical significance in separate regressions ($p = 0.029$ vs. 0.064, one-sided tests), indicating a difference in inference rather than effect size.

concerns alone do not drive belief adjustments. Although women consistently report higher (worse) ranks than men, I find no significant gender-treatment interactions, providing no support for **H6** (gender differences in social-image sensitivity) nor for **H8** (gender differences in self-image sensitivity). This indicates that while exposure influences belief reporting, its effect appears similar across genders, with no evidence of differential sensitivity to anticipated social costs—or to the (statistically insignificant) ego-related costs.

Result 2: Effects on self-placement accuracy. Building on the observed shifts in modal rank under anticipated exposure, I now examine whether social observability affects agents' self-confidence accuracy in placement—that is, the difference between their actual rank and the modal rank in their belief distribution. One possible concern is that these changes simply reflect performance differences. If agents who anticipate visibility or private feedback also perform worse under pressure, rank placement might shift mechanically.[55] However, as shown in Table G.1 (Appendix G), neither private feedback, public observability, nor their combination significantly affect real-effort task scores. Performance differences are therefore unlikely to explain the placement patterns observed in Round 1.

Table 4 presents OLS estimates of placement accuracy in Round 1, where positive values indicate overplacement and negative values underplacement. The models control for actual rank and academic background variables. Across all treatments, women place themselves significantly lower (i.e., worse) than men relative to the control condition: a difference of approximately 1.3 rank steps (all significant at the 1% level), corresponding to 0.22–0.25 standard deviations.[56]

Public observability (in the *Public* condition) also reduces overplacement relative to the *Control*. Treated participants lower their stated rank by approximately 1.6 steps—or about 0.28 standard deviations—(p between 0.02 and 0.045; see cols. (3)–(4) in Table 4), consistent with a cautious adjustment in self-assessment when belief accuracy is visible to others. In contrast, neither the *Private* nor the *Joint* condition yields a significant shift in placement. Importantly, gender-treatment interactions are not significant. Both men and women respond similarly to public exposure, though the baseline gender gap in placement persists. This suggests that social exposure shapes behaviour similarly across genders but does not eliminate underlying self-confidence asymmetries. These patterns are consistent with strategic self-presentation responses documented elsewhere and may have implications for settings where performance beliefs are made public (e.g., hiring or peer evaluation).

[55] In Schram et al. (2018) the authors find that men but not women, increase their number of attempted problems and competitive performance when anticipating that their result will be "rankable" by an observer *ex post*.

[56] Across treatments, women's average overplacement is lower by 1.41 ranks in the *Private*, 1.24 ranks in the *Joint*, and 1.25 ranks in the *Public* condition. Using the *Control* group SD of 5.75, this corresponds to gender gaps of 0.25, 0.22, and 0.22 standard deviations, respectively.

Table 4: Agents' placement, first round

	Private		Public		Joint	
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-1.41**	-0.52	-1.25**	-1.88	-1.24**	-1.46
	(0.51)	(1.22)	(0.41)	(1.56)	(0.39)	(1.00)
Treatment	-0.24	-0.14	-1.61*	-1.62*	-0.81	-0.85
	(0.48)	(0.55)	(0.65)	(0.78)	(0.49)	(0.49)
Treatment × Female		-0.19		0.10		0.06
		(0.85)		(0.84)		(0.85)
Round rank	✓	✓	✓	✓	✓	✓
Academic cont.	✓	✓	✓	✓	✓	✓
Female × Academic cont.		✓		✓		✓
Mean of dep. variable	3.11	3.11	2.28	2.28	2.57	2.57
Observations	274	274	275	275	277	277

Notes: This table reports OLS estimates of placement accuracy—defined as the difference between the modal belief rank and the agent's true rank. Each column pair compares one treatment condition to the *Control*. Regressors include treatment dummies, a gender indicator, and their interaction (in even-numbered models). All models control for realised rank and academic background (years of finished years of higher education and major field of study), with gender interactions included where specified. Standard errors in parentheses clustered at the session level. *** $p < 0.05$, *** $p < 0.01$, *** $p < 0.001$.

6.1.3 Mechanisms: Strategy or Shift in Beliefs

In addition to examining shifts in modal rank, I explore whether anticipation of public exposure induces broader changes in belief distributions. Two psychological channels are plausible: (1) avoidance of difficult emotions, such as reputational embarrassment, which would affect only the visible rank (the modal guess); or (2) a genuine shift in self-confidence, which would affect the full underlying belief distribution. If the former dominates, adjustments should be concentrated in the selected modal rank.[57] If the latter, the allocation as a whole should shift upward—that is, toward worse-ranked positions. To distinguish between these channels, I compare the distribution of non-modal rank guesses across the *Control* and *Public* conditions (see Figure 8). I develop two metrics: (i) a skewness measure capturing the asymmetry of rank weights around the modal guess; and (ii) the average minimum and maximum ranks selected (excluding the mode), reflecting the breadth and positioning of belief support.

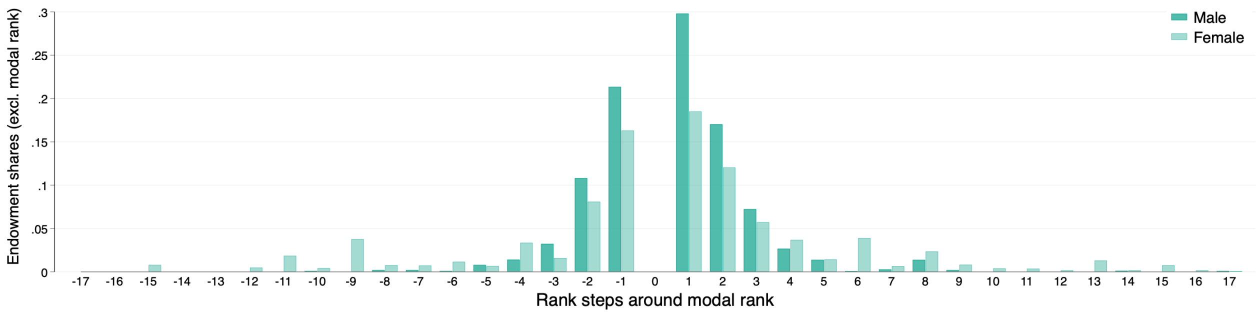
Results show that while, relative to the modal rank guess, skewness in the *Public* condition is slightly less positive than in the *Control*, the difference is modest and only marginally significant ($p = 0.09$, one-sided t-test). However, for the second measure, the minimum and maximum selected ranks shift meaningfully upward—towards worse ranks—under public exposure: the left tail (i.e., better ranks) increases by 0.78 rank steps, and the right tail (i.e., worse ranks) by 1.19 steps ($p < 0.01$ and $p = 0.017$, respectively).[58] Figure 8 visualises the average allocation of belief weights across non-modal ranks,

[57]This would imply a leftward extension of the belief distribution relative to the modal guess—more weight placed on better-than-modal ranks. While this suggests a longer left tail in intuitive terms, note that statistical skewness is measured relative to the mean, not the mode. Therefore, a modal shift alone does not necessarily change skewness unless the tails shift asymmetrically.

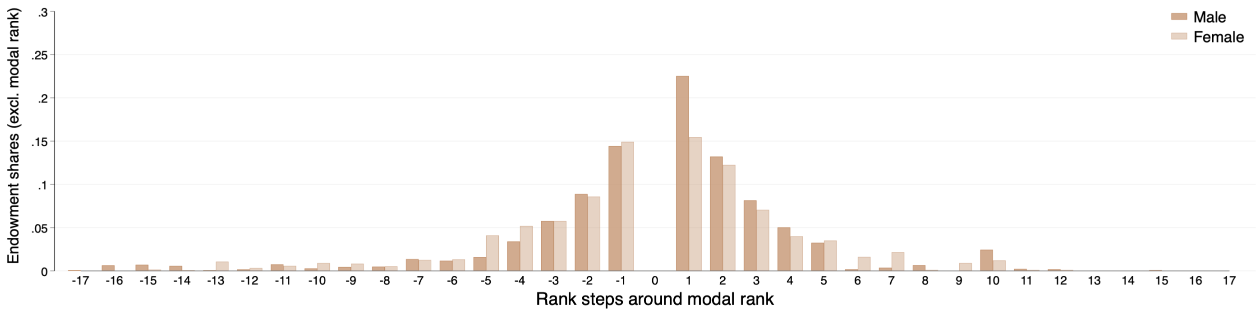
[58]Interestingly, the left-tail shift is driven by men; the right-tail shift by women.

Figure 8: Non-modal rank allocations, by gender—first round

(a) Control condition



(b) Public condition



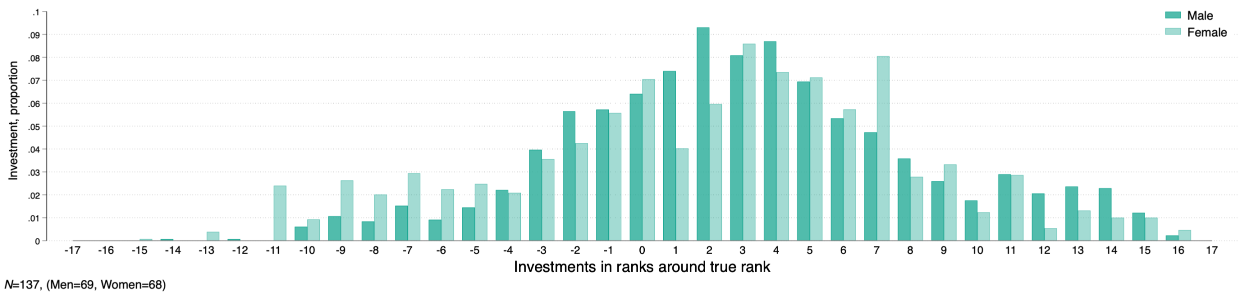
Notes: The figure displays the distribution of belief allocations across ranks in the first round, measured in rank-steps from the modal guess (the rank to which the most ECUs were allocated). Shares are computed from the remaining endowment after excluding the modal rank and reweighted to sum to 1. The top panel (Figure 8a) shows the *Control* condition; the bottom panel (Figure 8b) the *Public* condition. Distributions are shown separately by gender, with rightmost bars for females (lighter shading).

reweighted accordingly, by gender. The top panel (8a) shows the *Control* condition; the bottom panel (8b), the *Public* condition. Together, these findings suggest that public observability induces a broader recalibration of belief distributions—not merely a cosmetic adjustment of the modal-"visible"-rank. This points to a confidence-based mechanism rather than a purely strategic signalling response.

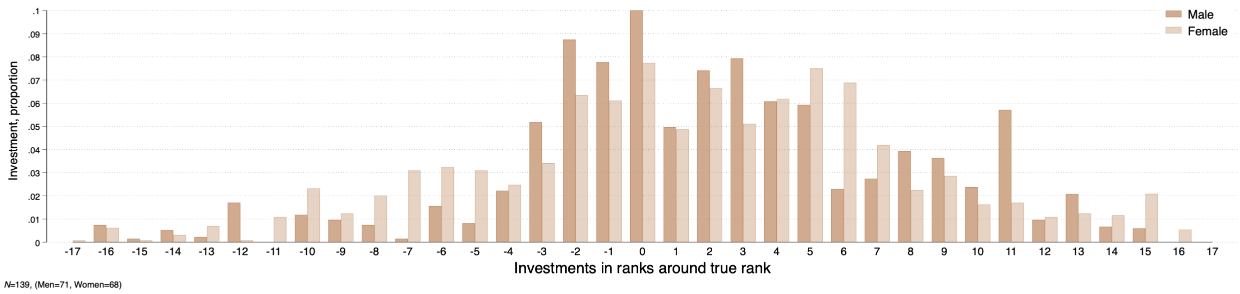
Beyond the modal rank: evidence from full belief allocations. Beyond modal beliefs, agents allocate ECUs across multiple ranks. While only about 3% of participants allocate their entire endowment to a single rank, most distribute it across a range of plausible outcomes. Although payoffs depend solely on correctly identifying the true rank, the design allows for risk-hedging—both in monetary terms and in terms of anticipated *social risk* when exposure is possible (*Public* and *Joint* treatments). Because only the modal rank is revealed to the principal, agents also have room for strategic self-misrepresentation: expressing confidence in one rank while hedging across others.

Figure 9: Rank Deviations: all ranks, by gender - first round

(a) Control condition



(b) Public condition



NOTES: The figure shows all rank allocations in self-confidence steps from the true ranks. Bars represent shares of the portfolio, by gender. The upper Figure 9a, displays allocations for the *Control* condition, and the lower Figure 9b for the *Public* condition. Treatments are separated by gender, showing females with lower colour-intensity compared to males.

Figure 9 visualises the distribution of allocations shares across all ranks, centred around each agent's true rank. Panel (a) shows the *Control* condition; Panel (b) shows the *Public* condition. Bars indicate the share of remaining ECUs (excluding the modal rank) allocated to ranks at different distances from the true rank, with negative values indicating underplacement. In the *Public* condition (Fig. 9b), both men and women shift their distributions leftward relative to the *Control*. For example, at a deviation of -2 (i.e., two ranks worse than the true rank), men allocate 8.7% of their remaining belief and women 6.3%. This supports the idea that agents shift belief mass away from the true rank when facing exposure—consistent with the behavioural adjustments described earlier.

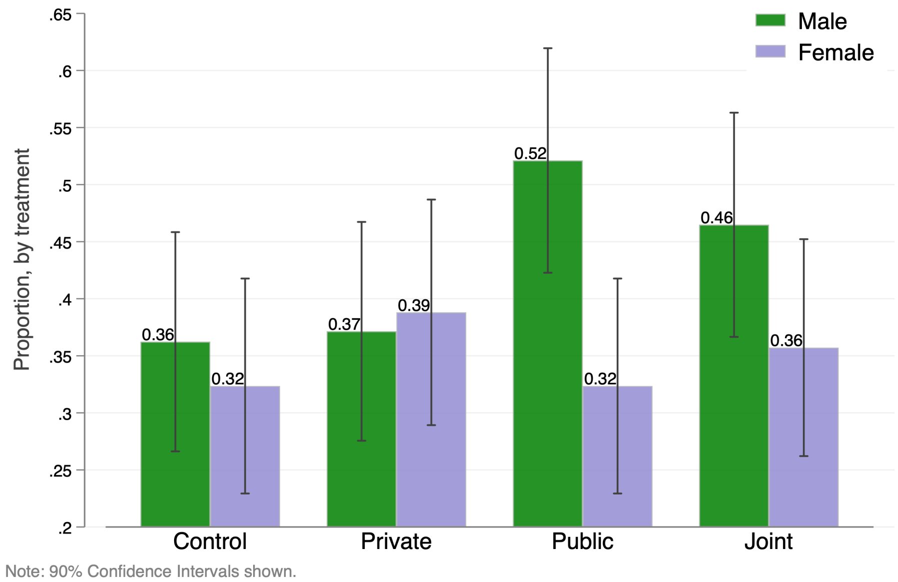
Having shown that public observability shapes reported confidence across the belief distribution, I now turn to its implications for payoff-relevant accuracy—specifically, whether agents assign any belief to their actual rank.

6.1.4 Consequences for accuracy and earnings (Round 1)

This subsection turns to the outcome-level implications of agents' rank assessments. Specifically, it examines whether agents allocate any belief weight to their true rank—a necessary condition for receiving earnings—and whether this likelihood differs by gender. The analysis provides a behavioural measure of accuracy that complements the directional self-assessment biases explored in earlier sections. This subsection tests Hypothesis H4.

Earnings outcomes: gender differences in belief-related earnings. Previously, I established that anticipation of exposing self-accuracy shift how modal rank beliefs are reported, as agents manage the risk of public overplacement. Agents required to disclose both their true and self-assessed rank to the principal tend to allocate more weight to ranks below their actual position. This suggests that such adjustments may also influence payoffs through broader changes in rank beliefs. To assess whether these belief shifts translate into more accurate rank allocations, we examine the share of agents allocating resources to their true rank. Figure 10 illustrates that while overall allotments to the true rank do not differ significantly across treatments, anticipated public belief exposure benefits men but not women. Under the *Public* and *Joint* treatments, men are substantially more likely to pick their true rank compared to men in the *Control* group, whereas women's allocation patterns remain unchanged.

Figure 10: Share of endowment allocated to true rank, by treatment and by gender



NOTES: The figure displays the share of each agent's endowment allocated to their true rank in Round 1, broken down by treatment and gender. Bars represent mean allocations, with the error bars indicating 90% confidence intervals.

These findings highlight a striking gender asymmetry in how social exposure affects self-assessment accuracy and payoffs. In the *Public* condition, men are 44% more likely to select their true rank (a 15.9 percentage point increase, $p = 0.06$), and in the *Joint* condition, 28% more likely (10.2 percentage points, $p = 0.22$), compared to the Control, although neither difference reaches conventional significance levels. Notably, the *Private* treatment–where feedback is received privately–does not generate any meaningful change for men, reinforcing that the observed effects are likely driven by social exposure rather than anticipation of self-image concerns and private updating.[59] However, Figure 10 shows that women's choices do not shift in response to public visibility. As a result, men in the *Public* condition are significantly more likely to receive any payoff compared to women, with a 19.8 percentage point gender gap in realised earnings (52% vs. 32%, $p = 0.018$ from a $\tilde{\chi}^2$-test).

[59] Comparing the *Private* and the *Control* condition (*diff.* $= 0.01$ percentage points, $p = 0.91$ from a $\tilde{\chi}^2$-test).

Table 5 confirms these results in a regression framework, estimating the likelihood of selecting the true rank in the first round, across treatments. Expressed as odds ratios, the estimates indicate that men in the *Public* condition are between 1.82 to 2.28 times more likely to pick their true rank, a substantial and significant effect. No comparable effect emerges for women, reinforcing the idea that public observability enhances men's self-assessment accuracy while leaving women's unchanged.

Table 5: Allocation to the true rank—first round

	Private		Public		Joint	
	(1)	(2)	(3)	(4)	(5)	(6)
ALLOCATION TO THE TRUE RANK						
Female	0.99	0.79	0.62	0.79	0.82	0.79
	(0.29)	(0.32)	(0.18)	(0.31)	(0.24)	(0.32)
Treatment	1.33	1.04	1.82*	2.28*	1.31	1.26
	(0.40)	(0.45)	(0.52)	(0.82)	(0.39)	(0.49)
Treatment × Female		1.61		0.62		1.09
		(0.95)		(0.35)		(0.65)
Round score		✓	✓	✓	✓	✓
Academic cont.		✓	✓	✓	✓	✓
Mean of dep. variable		.361	.385	.385	.375	.375
Observations	274	274	275	275	277	277

NOTES: The table shows logistic regression results expressed as odds-ratios, estimating the likelihood of selecting the *true rank* as a function of treatment conditions, gender, and their interaction. Each column-pair represents a different treatment condition compared to the *Control*. In Column 1 and 2 is the Private treatment, in 3 and 4 the Public treatment and in 5 and 6 the Joint treatment condition. The models control for the real-effort task scores, years of completed education, and the number of selected ranks that agents selected. Robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

A more detailed decomposition of these allocation patterns—showing that the treatment effect for men is driven primarily by the extensive rather than the intensive margin—is reported in Appendix H.2.

In sum, that *Public* men allocating on average more to the true rank, compared to men in the *Control* condition, was primarily driven by the larger likelihood amongst those men to (at all) believing in the true rank, rather than differences in the belief-intensity, or changes in certainty. Anticipating public observability leads agents—particularly men—to lower their stated rank and more often select their true rank, improving their likelihood of earning. Women, who already start with lower self-placement, do not exhibit further adjustments that enhance their accuracy. These patterns highlight a gender asymmetry in the effects of social exposure: it may help men calibrate better while leaving women's lower baseline confidence unchanged.

6.2 Experiencing Exposure (Round 2)

This section first describes how performance, stated rank beliefs, and received signals change between rounds (6.2.1), followed by treatment effects on subsequent rank beliefs in response to prior signals (6.2.2). It then examines asymmetries in these responses and whether they differ by gender, and concludes by examining whether these patterns differ across feedback environments.

6.2.1 Changes in Performance and Belief Patterns (Round 2)

Round-2 descriptives show only modest changes relative to Round 1. Performance remains stable across treatments, with average RET scores increasing by roughly one point.[60] Overall, rank-allocation patterns are similar across treatments. Preferred ranks (modal beliefs) are more conservative in the *Private* and *Joint* conditions than in the *Control*, consistent with feedback-driven belief updating. Consequently, the placement accuracy exhibits the same qualitative pattern as in Round 1: underplacement is somewhat more common in the *Private* condition, accurate placement is most frequent in *Joint*, and overplacement is most pronounced in *Control* and *Public*.

Receiving feedback on Round-1 errors (*Private* and *Joint* conditions) does not appear to substantially alter belief precision. This holds whether precision is measured as the share of ECUs placed on the modal rank or as the number of ranks to which agents allocate any weight. Both measures vary little across treatments and show only small gender differences. The lowest panel, *Feedback Signals*, summarises the Round-1 accuracy information revealed to agents. Most agents receive a negative signal—that is, they learn that they overplaced themselves in Round 1—and these errors tend to be larger in magnitude than underplacement errors.[61] These descriptive patterns provide the empirical context for analysing how agents revise their stated ranks in Round 2.

[60] Average *Score change (R1–R2)* refers to the change in RET points since Round 1 and shows improvements of about one point (e.g. +1.21 in *Control* and +1.27 in *Private*).

[61] Across the two feedback treatments, the average absolute signal is approximately four rank positions out of the 17 possible misplacements, indicating that most agents discover a sizeable error in Round 1.

Table 6: Descriptive Statistics, by treatment and gender—second round

		Control				Private				Public				Joint				Pairwise comp.		
	Scale	All (1)	Male (2)	Female (3)	p-val. (4)	All (5)	Male (6)	Female (7)	p-val. (8)	All (9)	Male (10)	Female (11)	p-val. (12)	All (13)	Male (14)	Female (15)	p-val. (16)	p-val. (1)-(5)	p-val. (1)-(9)	p-val. (1)-(13)
REAL-EFFORT TASK PERFORMANCE																				
Score	7–30	17.91	17.94	17.88	0.918	17.54	17.64	17.44	0.765	17.21	17.68	16.75	0.123	18.04	18.65	17.46	0.061	0.373	0.081	0.760
		(3.09)	(3.28)	(2.92)		(3.75)	(4.63)	(2.68)		(3.49)	(3.86)	(3.05)		(3.66)	(3.78)	(3.47)				
Score change (R2-R1)	-4–9	1.21	1.20	1.22	0.952	1.27	1.39	1.15	0.431	0.86	1.05	0.68	0.293	1.07	1.06	1.07	0.975	0.790	0.149	0.561
		(1.95)	(2.05)	(1.86)		(1.81)	(1.85)	(1.77)		(2.00)	(2.12)	(1.88)		(2.00)	(1.97)	(2.05)				
RANK ALLOCATIONS																				
No. of ranks	1–18	5.07	4.95	5.19	0.631	4.93	4.98	4.88	0.849	5.29	5.79	4.81	0.084	4.56	4.31	4.80	0.239	0.696	0.564	0.108
		(2.82)	(2.30)	(3.25)		(3.09)	(3.13)	(3.08)		(3.28)	(3.55)	(2.95)		(2.40)	(2.21)	(2.55)				
Only one	0/1	0.03	0.03	0.03	0.964	0.06	0.03	0.09	0.157	0.02	0.02	0.03	0.586	0.02	0.03	0.01	0.524	0.232	0.702	0.709
Preferred rank	1–18	5.50	4.74	6.22	0.044	7.88	7.00	8.74	0.038	6.64	5.36	7.86	0.001	7.75	6.75	8.68	0.022	0.000	0.031	0.000
		(4.27)	(3.72)	(4.64)		(4.86)	(5.01)	(4.59)		(4.36)	(4.03)	(4.35)		(4.91)	(4.83)	(4.83)				
Share allocated[1]	0–1	0.37	0.36	0.38	0.526	0.39	0.37	0.41	0.190	0.36	0.33	0.40	0.011	0.39	0.39	0.38	0.694	0.441	0.673	0.497
		(0.17)	(0.17)	(0.17)		(0.20)	(0.16)	(0.22)		(0.15)	(0.14)	(0.15)		(0.15)	(0.16)	(0.15)				
Placement[2]	-16–17	2.60	3.33	1.90	0.136	0.80	1.42	0.19	0.081	1.58	2.36	0.83	0.129	0.57	0.69	0.46	0.748	0.003	0.144	0.001
		(5.58)	(4.64)	(6.31)		(4.09)	(3.93)	(4.18)		(5.87)	(5.56)	(6.11)		(4.10)	(3.25)	(4.79)				
Underplac.	0/1	0.24	0.18	0.30	0.098	0.37	0.29	0.44	0.065	0.33	0.29	0.38	0.273	0.29	0.26	0.32	0.466	0.031	0.107	0.388
Accurate plac.	0/1	0.05	0.08	0.03	0.221	0.16	0.20	0.12	0.207	0.08	0.11	0.06	0.307	0.18	0.23	0.13	0.130	0.005	0.329	0.001
Overplac.	0/1	0.70	0.74	0.67	0.339	0.48	0.52	0.44	0.395	0.59	0.61	0.57	0.633	0.53	0.51	0.55	0.621	0.000	0.042	0.003
FEEDBACK SIGNALS																				
Signal steps (abs.)	0–16					4.90	5.67	4.15	0.036					4.66	4.45	4.87	0.505			
						(4.21)	(4.46)	(3.84)						(3.66)	(3.33)	(3.95)				
Underpl. feedback	0/1					0.26	0.23	0.29	0.379					0.37	0.37	0.36	0.934			
Signal steps (abs.)	0–14					0.89	0.82	0.96	0.720					1.25	1.28	1.22	0.874			
						(2.21)	(2.20)	(2.24)						(2.16)	(2.24)	(2.10)				
Acc. feedback	0/1					0.12	0.08	0.16	0.125					0.06	0.06	0.06	0.931			
Overpl. feedback	0/1					0.62	0.70	0.54	0.068					0.57	0.57	0.58	0.902			
Signal steps (abs.)	0–16					4.01	4.85	3.19	0.032					3.42	3.17	3.65	0.504			
						(4.47)	(4.81)	(3.99)						(4.16)	(3.78)	(4.50)				
N		135	66	69	135	134	66	68	134	135	66	69	135	134	65	69	134	269	270	269

[1] Agents' endowment allocations to their most preferred rank must be ≤ 2 ECUs, setting the lower range of approximately 0.11 (from 2/19).

[2] Placement is measured as the true rank minus the preferred rank, so that a negative (positive) number implies the underplacement (overplacement) of agent's rank and a zero signifies an accurate placement of the most preferred rank.

NOTES: This table reports descriptive statistics from the second round, by treatment and gender. The upper panels summarise real-effort task (RET) performance and rank-allocation behaviour in Round 2, including the number of ranks selected, the preferred rank, the share allocated to the preferred rank, and measures of placement accuracy. The lower panel summarises feedback signals received after Round 1 for treatments in which belief accuracy is revealed to agents (Private and Joint). Feedback indicators classify signals as negative (overplacement), accurate, or positive (underplacement), and corresponding absolute signal magnitudes capture the size of the performance discrepancy revealed. Columns report means separately for all agents (col. 1), men (col. 2), and women (col. 3), with p-values in col. 4 based on t-tests for continuous variables or χ^2-tests for binary variables.

6.2.2 Treatment Effects (Round 2)

After experiencing the exposure conditions from Round 1, agents adjust their subsequent rank beliefs in the direction implied by the feedback signal: they revise their stated rank *downwards* (towards a worse believed position) after receiving a *negative* signal of overplacement, and *upwards* (towards a better believed position) after a *positive* signal of underplacement. However, these adjustments are not symmetric; responses differ in magnitude across signal types. This asymmetry also differs by gender, with women adjusting more after overplacement and men adjusting more after underplacement.

Result 1: Effects on subsequent rank beliefs. To test whether agents incorporate past feedback into their updated rank beliefs, I estimate a linear model of belief revision in Round 2. The dependent variable is the change in modal rank between rounds 1 and 2 ($\Delta\widehat{r}_2 = \widehat{r}_1 - \widehat{r}_2$), where positive values indicate an upward revision (towards a better believed position) and negative values indicate a downward revision (towards a worse believed position). The main explanatory variable is the round-1 feedback signal, defined as the difference between an agent's modal guess and her realised rank ($\text{signal}_1 = \widehat{r}_1 - r_1^*$). A negative signal indicates overplacement (the agent performed worse than she believed and should revise towards a worse, higher-numbered rank), while a positive signal indicates underplacement (she performed better than she believed and should revise towards a better, lower-numbered rank). The model includes two additional predictors. First, the change in real-effort task performance between rounds (ΔRET), capturing new internal information that could influence beliefs. Second, a dummy for the *Joint* condition (with the *Private* condition as the reference), to test whether belief updating differs when feedback is publicly visible. The sample is limited to the feedback treatments, in which agents actually did receive signals.

Results in Table 7 show that agents significantly revise their beliefs in response to feedback. The coefficient on the signal is 0.70 ($p < 0.001$), suggesting that agents incorporate roughly 70% of the round-1 mismatch into their round-2 rank belief. This is strong evidence for Hypothesis **H9**, which predicts directional updating in response to prior signals. Agents also incorporate their own changes in performance: the coefficient on ΔRET is 0.26 ($p = 0.025$), indicating that belief revisions also respond to recent task outcomes. While smaller than the feedback coefficient, this implies that agents weigh both external feedback and internal performance cues when updating. However, because rank is relative, performance changes only influence beliefs when agents perceive their own change to differ from others in their session. To address potential floor or ceiling effects, Column (2) includes actual round-2 rank as a control. The results remain stable, indicating that the main effects are not mechanically driven by agents near the boundaries of the rank distribution.

Taken together, these results provide strong support for Hypothesis **H9**: agents partially—but systematically—incorporate prior feedback when updating their self-assessed rank, consistent with learning from performance-based signals.

Result 2: Asymmetries in responses to prior rank signals. To test whether agents respond differently to feedback depending on its valence, I estimate a specification that interacts the feedback signal with indicators for overplacement and underplacement feedback, omitting cases of accurate

Table 7: Belief updating: response to feedback and performance signals (Round 2)

	Baseline updating		Asymetric updating	
	(1)	(2)	(3)	(4)
Feedback signal (t-1)	0.70***	0.77***		
	(0.037)	(0.049)		
Underpl. signal (t-1)			0.91***	0.93***
			(0.144)	(0.142)
Overpl. signal (t-1)			0.60***	0.69***
			(0.060)	(0.072)
Δ RET score	0.26*	0.41**	0.23	0.39**
	(0.108)	(0.126)	(0.127)	(0.128)
Feedback × Public info. (Joint)	-0.26	-0.32	-0.23	-0.27
	(0.390)	(0.391)	(0.396)	(0.397)
Round rank		✓		✓
Academic covariates		✓	✓	✓
Mean of dep. variable	-1.75	-1.77	-1.92	-1.92
Observations	261	260	239	239

NOTES: This table reports OLS estimates of belief updating in Round 2. The dependent variable is the change in modal rank ($\Delta\widehat{r}_{i,2} = \widehat{r}_{i,1} - \widehat{r}_{i,2}$). A positive value indicates that the agent revised her belief upwards (towards a better, lower-numbered rank), while a negative value indicates a downward revision (towards a worse, higher-numbered rank). The key predictor is the round-1 feedback signal ($\text{signal}_{i,1} = \widehat{r}_{i,1} - r^*_{i,1}$), where a negative value indicates overplacement and a positive value indicates underplacement. Columns (3)–(4) split the signal into separate effects for underplacement and overplacement. All models include performance change (ΔRET) and a dummy for public observability (*Joint*). Columns (2) and (4) additionally control for round-2 rank and academic background. Standard errors are in parentheses, clustered at the session level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

feedback.[62] The outcome is belief updating in Round 2 ($\Delta\widehat{r}_2 = \widehat{r}_1 - \widehat{r}_2$), and the two key predictors capture the amount of revision per rank of feedback discrepancy. That is, the coefficient on *Underpl. signal (t-1)* indicates how strongly agents respond when told they underplaced themselves (i.e., they did better than they thought), and the coefficient on *Overpl. signal (t-1)* captures response to overplacement feedback (i.e., they did worse than they thought). No intercept terms are included for feedback types; each coefficient directly represents the updating slope within that feedback category.

As shown in Table 7 (columns 3–4), both coefficients are statistically significant, confirming that agents update beliefs in the direction of the signal. However, the magnitude differs: in response to underplacement feedback, agents revise towards a better (lower-numbered) believed rank by 0.91 rank positions per unit of signal ($p < 0.001$), while the response to overplacement feedback is smaller at 0.60 rank positions per signal unit ($p < 0.001$). This asymmetry suggests that agents are more willing to incorporate flattering feedback than critical feedback into their belief revisions. While not conclusive, this pattern is consistent with motivated reasoning and positivity bias in self-assessment.

The difference in slope magnitudes provides strong evidence in support of Hypothesis **H10**, which predicts that feedback-based updating is asymmetric, with agents more responsive to signals that

[62]This excludes 21 observations where modal guess exactly matched true rank in Round 1.

affirm rather than threaten their prior self-view.

Result 3: Gender differences in responses to prior rank signals. To test H11—that women and men respond differently to feedback depending on its direction—I extend the analysis of H10 using a proportional adjustment variable (*adjust_r2*). This measure captures how much of the prior feedback signal (i.e., the difference between the agent's modal rank belief and her true performance rank in the previous round) is reflected in her updated *modal rank belief* in the subsequent round. A value of 1 indicates that the agent fully adjusted her modal belief to match the feedback signal; 0 means no adjustment; values greater than 1 reflect over-adjustment, and negative values indicate belief revision in the wrong direction.[63]

Aggregating across *feedback signal types* (i.e., pooling over signal valence), agents respond more strongly to overplacement feedback (mean = 0.80) than to underplacement feedback (mean = 0.55), with the difference marginally significant ($p = 0.058$, one-sided t-test). This replicates the asymmetry predicted in H10 using a more interpretable measure of belief revision.

When disaggregated by gender, clear differences emerge. Among men, the asymmetry reverses: they incorporate more feedback following underplacement (mean = 0.83) than overplacement (mean = 0.61), with a one-sided p-value of 0.050. Among women, the asymmetry aligns with the pooled result: adjustment is significantly greater after overplacement (mean = 1.02) than after underplacement (mean = 0.33), with a highly significant difference ($p = 0.008$). These opposing patterns suggest that the overall asymmetry in feedback responsiveness (as captured in H10) conceals gender-specific dynamics. Formal comparisons of gender gaps within each feedback type confirm this divergence. Women adjust significantly more than men after overplacement ($p = 0.035$), whereas men adjust significantly more than women after underplacement ($p = 0.015$). These findings provide strong support for **H11**.

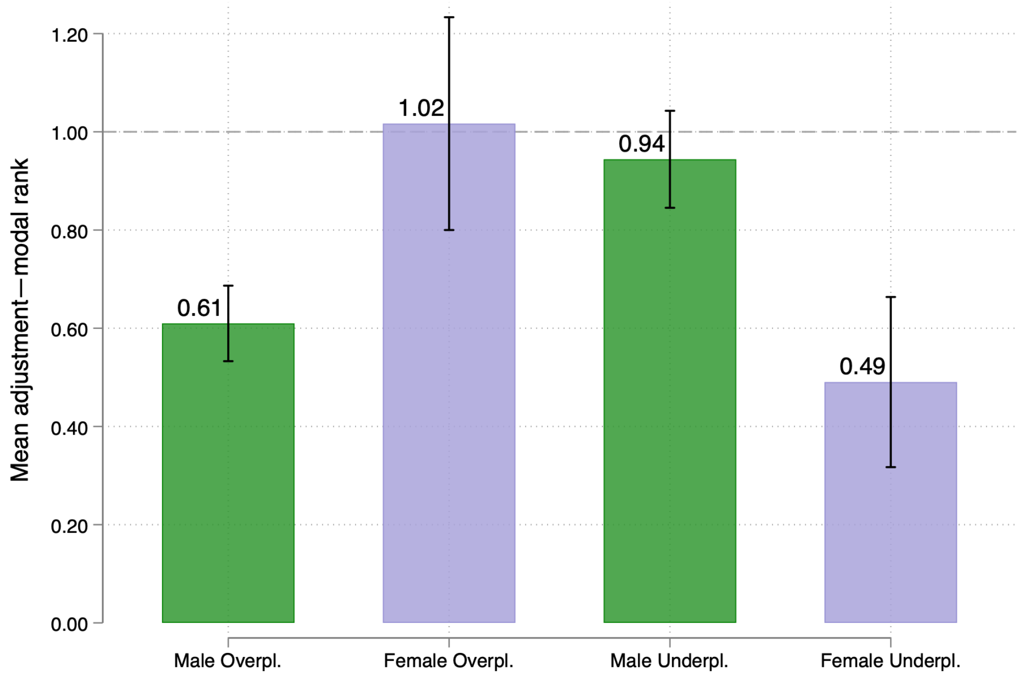
The results remain directionally robust when controlling for education and performance: OLS regressions confirm the magnitude of gender differences, although the overplacement effect is only marginally significant at conventional levels ($p = 0.074$).[64] Given sample size constraints and heterogeneity in adjustment behaviour, the regression estimates serve primarily as a robustness check.

Together, these findings align closely with the conceptual model (Section 3), in which gender-specific responsiveness to negative and positive signals is captured by asymmetric updating parameters η^+ and η^-. Importantly, no gender difference appears when pooling across feedback types, underscoring the need to disaggregate feedback by signal valence when analysing belief formation dynamics.

[63] In cases of exact accuracy in Round 1, the adjustment is set to 1 if no revision is made and -1 if any revision occurs; see Appendix E.1 for details and the formal definition.

[64] See Appendix I Table I.1 for full regression results.

Figure 11: Proportional adjustment in Round 2 modal rank beliefs, by gender and type of feedback signal



NOTES: The bars show the mean adjustment in modal rank beliefs *adjust_r2*, from Round 1 to Round 2, conditional on signal direction (overplacement vs. underplacement) and gender. A value of 1 indicates full correction of the Round 1 misplacement error. Standard errors are shown as capped vertical lines. The sample is restricted to agents in the *Private* and *Joint* conditions, who received feedback on their Round 1 accuracy.

Result 4: Effects of public observability on responses to prior rank signals. Finally, although the coefficient on public observability in Table 7 suggests slightly weaker updating in the *Joint* condition compared with the *Private* condition, two-sample t-tests indicate that these differences are not statistically significant. There is some exploratory indication that men become more responsive to feedback when both private feedback and public observability are present—hinting at a possible accountability effect—but this pattern is not robust across signal types. Accordingly, I find no systematic evidence in support of **H12**.

The experience-driven results confirm that belief updating is systematic, directionally sensitive, and gender-differentiated. Agents incorporate feedback signals into their future self-assessments, but do so asymmetrically depending on the type of information and their own characteristics. While overplacement prompts stronger corrections on average, men and women respond differently depending on whether the feedback affirms or threatens their prior beliefs. These dynamics suggest that even when feedback is structured and evenly distributed, downstream belief trajectories—and potentially confidence gaps—can diverge meaningfully. The findings highlight that belief updating is not simply mechanical, but instead reflects heterogeneous psychological sensitivities to error and exposure, as captured in the conceptual model. These insights motivate broader implications for feedback design, public observability, and institutional mechanisms aiming to mitigate persistent gaps in self-assessment and advancement.

7 Conclusions

This paper examines how individuals form and revise beliefs about their relative performance in competitive settings, and how social and psychological dimensions of these environments shape that process. While existing work often treats competitions as arenas that reveal underlying traits such as confidence or willingness to compete, this paper reverses that perspective: it proposes that competitive environments are inherently informational, and can actively shape individuals' self-assessments through social exposure and feedback. Building on a stylised model of belief management under observability, I show that individuals weigh not only accuracy incentives but also perceived reputational risks when reporting beliefs about their relative standing. The model predicts that belief distortion arises from both self-image and social-image concerns, and that exposure to belief error—whether private or public—can generate asymmetric behavioural responses. These predictions are tested in a controlled laboratory experiment designed to isolate the informational structure of competitive interactions across repeated rounds.

The experiment reveals several robust empirical patterns. At baseline, women report more conservative self-assessments than men, even when performance is held constant. This gender gap persists across treatments, consistent with prior evidence on gendered self-beliefs. However, when agents anticipate that their belief accuracy will be publicly observable, both men and women reduce their overplacement substantially, suggesting that social exposure powerfully constrains inflated self-assessments. Notably, the effect of observability is additive to the baseline gender gap, leading to especially low self-assessed ranks among women under public conditions. These results confirm the role of audience structure in shaping strategic self-presentation, and underscore the relevance of social-image motives even when no material consequences are at stake.

In the second phase of the experiment, I analyse how agents adjust their beliefs dynamically in response to personalised feedback. These results show that belief updating is systematic but asymmetric: agents incorporate signals of prior overplacement more strongly than signals of underplacement. When disaggregated by gender, this asymmetry reflects opposing patterns. Women respond more strongly to negative feedback about having overestimated their rank, while men respond more strongly to signals suggesting modesty or underplacement. These results suggest that belief dynamics are not only error-sensitive but also group-sensitive—shaped by differential tolerances for self-exposure and correction. The asymmetries align closely with the model's structure, where updating weights differ by signal valence and agent type.

At the same time, several limitations temper these findings. The experiment abstracts from many features of real-world competitions: social exposure is non-personalised, interaction is passive, and reputational consequences are contained. This likely places the estimates of social feedback effects at the lower end of what might be observed in real organisational settings. Moreover, while the laboratory allows for clean identification, it cannot capture how feedback and belief calibration interact with long-term institutional sorting, power dynamics, or the reputational capital accumulated in repeated high-stakes settings. Nevertheless, the experimental setting is especially well suited to disentangling psychological mechanisms from pecuniary incentives—precisely because social and monetary motives are difficult to separate in real-world environments. By controlling exposure, feedback, and incentives, the design isolates core belief formation dynamics that are often hidden in observational data. The results contribute to a growing literature on motivated beliefs, gender and competitiveness, and the behavioural foundations of inequality. They point to the importance of designing evaluative environments that are sensitive to asymmetric reactions to feedback—especially when public visibility and performance assessment intersect.

By shifting attention from static traits to contextually triggered reactions, this study contributes to a deeper understanding of how competitive environments shape self-beliefs—and how institutions may inadvertently amplify or mitigate confidence gaps through their informational design. Taken together, this paper shows that belief formation in competitive settings is shaped by more than ability or information—it is structured by visibility, psychology, and social expectation. Understanding these mechanisms is essential for designing institutions that aim to evaluate individuals fairly, encourage confidence, and mitigate unintended asymmetries in advancement and representation.

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753.

Alamaa, C. (2024). The employers' judgement - evaluating (biased) rank information in hiring decisions: A lab-experiment. *Mimeo, forthcoming.*

Alan, S. and Ertac, S. (2019). Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association*, 17(4):1147–1185.

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., and Tungodden, B. (2016). Willingness to compete: Family matters. *Management Science*, 62(8):2149–2162.

Apicella, C. L., Demiral, E. E., and Mollerstrom, J. (2017). No gender difference in willingness to compete when competing against self. *American Economic Review*, 107(5):136–40.

Armantier, O. and Treich, N. (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, 62:17–40.

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour economics*, 30:32–40.

Ball, S., Eckel, C., Grossman, P. J., and Zame, W. (2001). Status in markets. *The Quarterly Journal of Economics*, 116(1):161–188.

Ball, S. and Eckel, C. C. (1998). The economic value of status. *The Journal of socio-economics*, 27(4):495–497.

Baumeister, R. F. (1999). The nature and structure of the self: An overview. *The Handbook of the Self*, pages 1–20.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Benbow, C. P. and Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science*, 222(4627):1029–1031.

Berlin, N. and Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, 130:320–336.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865.

Bodner, R. and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, 1(105):26.

Brandts, J., Gërxhani, K., and Schram, A. (2020). Are there gender differences in status-ranking aversion? *Journal of Behavioral and Experimental Economics*, 84:101485.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, 62(12):3439–3449.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3):1409–1447.

Buser, T., Niederle, M., and Oosterbeek, H. (2021a). Gender differences in tournament choices: Risk preferences versus overconfidence. *Journal of the European Economic Association*, 20(4):1595–1631.

Buser, T., Niederle, M., and Oosterbeek, H. (2024). Can competitiveness predict education and labor market outcomes? evidence from incentivized choice and survey measures. *Review of Economics and Statistics*, pages 1–45.

Buser, T., Ranehill, E., and Van Veldhuizen, R. (2021b). Gender differences in willingness to compete: The role of public observability. *Journal of Economic Psychology*, 83:102366.

Buser, T. and Yuan, H. (2019). Do women give up competing more easily? evidence from the lab and the dutch math olympiad. *American Economic Journal: Applied Economics*, 11(3):225–52.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Cassar, A. and Rigdon, M. L. (2021a). Option to cooperate increases women's competitiveness and closes the gender gap. *Evolution and Human Behavior*, 42(6):556–572.

Cassar, A. and Rigdon, M. L. (2021b). Prosocial option increases women's entry into competition. *Proceedings of the National Academy of Sciences*, 118(45):e2111943118.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chow, C. W. (1983). 1983 competitive manuscript award: The effects of job standard tightness and compensation scheme on performance: An exploration of linkages. *Accounting Review*, pages 667–685.

Coffman, K., Ugalde Araya, M. P., and Zafar, B. (2024). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *Economic Inquiry*, 62(3):957–983.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660.

Cortés, P., Pan, J., Pilossoph, L., Reuben, E., and Zafar, B. (2023). Gender differences in job search and the earnings gap: Evidence from the field and lab. *The Quarterly Journal of Economics*, 138(4):2069–2126.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Dargnies, M.-P. (2012). Men too sometimes shy away from competition: The case of team competition. *Management Science*, 58(11):1982–2000.

De la Rosa, L. E. (2011). Overconfidence and moral hazard. *Games and Economic Behavior*, 73(2):429–451.

Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American economic review*, 101(2):556–590.

Dreber, A., Von Essen, E., and Ranehill, E. (2011). Outrunning the gender gap—boys and girls compete equally. *Experimental Economics*, 14:567–582.

Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, pages 247–296. Elsevier.

Duval, S. and Wicklund, R. A. (1973). Effects of objective self-awareness on attribution of causality. *Journal of experimental social Psychology*, 9(1):17–31.

Dweck, C. S. (1999). *Self-theories: Their Role in Motivation, Personality, and Development.* Psychology Press, New York.

Egerod, B., Staer, A., and Tranæs, T. (2022). Gender differences in self-promotion: Evidence from a field experiment. *European Economic Review*, 141:103958.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

Ellingsen, T. and Johannesson, M. (2007). Paying respect. *Journal of Economic Perspectives*, 21(4):135–150.

Else-Quest, N. M., Higgins, A., Allison, C., and Morton, L. C. (2012). Gender differences in self-conscious emotional experience: a meta-analysis. *Psychological bulletin*, 138(5):947.

Enke, B., Graeber, T., and Oprea, R. (2023). Confidence, self-selection, and bias in the aggregate. *American Economic Review*, 113(7):1933–1966.

Ewers, M. and Zimmermann, F. (2015). Image and misreporting. *Journal of the European Economic Association*, 13(2):363–380.

Exley, C. L. and Kessler, J. B. (2022). The gender gap in self-promotion*. *The Quarterly Journal of Economics*, 137(3):1345–1381.

Exley, C. L. and Kessler, J. B. (2023). Information avoidance and image concerns. *The Economic Journal*, 133(656):3153–3168.

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford University Press.

Flory, J. A., Leibbrandt, A., and List, J. A. (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1):122–155.

Fluchtmann, J., Glenny, A. M., Harmon, N. A., and Maibom, J. (2024). The gender application gap: Do men and women apply for the same jobs? *American Economic Journal: Economic Policy*, 16(2):182–219.

Frey, B. S. (2007). Awards as compensation. *European Management Review*, 4(1):6–14.

Gill, D. and Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, 5(2):351–376.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal* *of the Economic Science Association*, 1(1):114–125.

Grossman, Z. and van der Weele, J. (2017). Self-image and willful ignorance: A theoretical and experimental exploration. *Journal of the European Economic Association*, 15(1):173–217.

Gupta, N. D., Poulsen, A., and Villeval, M. C. (2005). Male and female competitive behavior-experimental evidence. *IZA Discussion Papers*, 1833.

Haeckl, S. (2022). Image concerns in ex-ante self-assessments–gender differences and behavioral consequences. *Labour Economics*, 76:102166.

Halko, M.-L. and Sääksvuori, L. (2017). Competitive behavior, stress, and gender. *Journal of Economic Behavior & Organization*, 141:96–109.

Harrison, G. W., Martínez-Correa, J., Swarthout, J. T., and Ulm, E. R. (2015). Eliciting subjective probability distributions with binary lotteries. *Economics Letters*, 127:68–71.

Healy, A. and Pate, J. (2011). Can teams help to close the gender competition gap? *The Economic Journal*, 121(555):1192–1204.

Heffetz, O. and Frank, R. H. (2011). Preferences for status: Evidence and economic implications. In *Handbook of social economics*, volume 1, pages 69–91. Elsevier.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, 80(3):984–1001.

Huffman, D., Raymond, C., and Shvets, J. (2022). Persistent overconfidence and biased memory: Evidence from managers. *American Economic Review*, 112(10):3141–3175.

Hyde, J. S. (2005). The gender similarities hypothesis. *American psychologist*, 60(6):581.

Kang, L., Lei, Z., Song, Y., and Zhang, P. (2024). Gender differences in reactions to failure in high-stakes competition: evidence from the national college entrance exam retakes. *Journal of Political Economy Microeconomics*, 2(2):355–397.

Kogelnik, M. (2022). Performance feedback and gender differences in persistence. *SSRN Electronic Journal*.

Kőszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature*, 52(4):1075–1118.

Kuhn, P. and Villeval, M. C. (2015). Are women more attracted to co-operation than men? *The Economic Journal*, 125(582):115–140.

Ludwig, S., Fellner-Röhling, G., and Thoma, C. (2017). Do women have more shame than men? an experiment on self-assessment and the shame of overestimating oneself. *European Economic Review*, 92:31–46.

Mijović-Prelec, D. and Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1538):227–240.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11):7793–7817.

Morton, H. and Dweck, C. (2003). Self-theories: Their role in motivation, personality, and development. *Psychological Review*, 110(1):93–110.

Niederle, M. (2016). Gender. In Kagel, J. H. and Roth, A. E., editors, *The Handbook of Experimental*

Economics, Volume 2, pages 481–562. Princeton University Press, Princeton, NJ.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Offerman, T., Sonnemans, J., Van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, 76(4):1461–1489.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of economic literature*, 37(1):7–63.

Reuben, E., Wiswall, M., and Zafar, B. (2017). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal*, 127(604):2153–2186.

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton University Press, Princeton, NJ.

Santos-Pinto, L. (2008). Positive self-image and incentives in organisations. *The Economic Journal*, 118(531):1315–1332.

Santos-Pinto, L. (2012). Labor market signaling and self-confidence: Wage compression and the gender pay gap. *Journal of Labor Economics*, 30(4):873–914.

Sarsons, H. and Xu, G. (2021). Confidence men? evidence on confidence and gender among top economists. In *AEA Papers and Proceedings*, volume 111, pages 65–68. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Sautmann, A. (2013). Contracts for agents with biased beliefs: Some theory and an experiment. *American Economic Journal: Microeconomics*, 5(3):124–156.

Schotter, A. and Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, 6(1):103–128.

Schram, A., Brandts, J., and Gërxhani, K. (2018). Social-status ranking: a hidden channel to gender inequality under competition. *Experimental Economics*, pages 1–23.

Schwardmann, P., Tripodi, E., and Van der Weele, J. J. (2022). Self-persuasion: Evidence from field experiments at international debating competitions. *American Economic Review*, 112(4):1118–1146.

Sedikides, C. and Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true. *Advances in Experimental Social Psychology*, 29:209–269.

Shastry, G. K. and Shurchkov, O. (2024). Reject or revise: Gender differences in persistence and publishing in economics. *Economic Inquiry*, 62(3):933–956.

Shastry, G. K., Shurchkov, O., and Xia, L. V. (2020). Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics*, 88:101592.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1):4–28.

Stigler, G. J. (1962). Information in the labor market. *Journal of political economy*, 70(5, Part 2):94–105.

Sutter, M. and Glätzle-Rützler, D. (2015). Gender differences in the willingness to compete emerge early in life and persist. *Management Science*, 61(10):2339–2354.

Tracy, J. L. and Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry*, 15(2):103–125.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650.

Van Veldhuizen, R. (2022). Gender differences in tournament choices: Risk preferences, overconfidence, or competitiveness? *Journal of the European Economic Association*, 20(4):1595–1618.

Weiss, Y. and Fershtman, C. (1998). Social status and economic performance:: A survey. *European Economic Review*, 42(3-5):801–820.

Wozniak, D., Harbaugh, W. T., and Mayr, U. (2016). The effect of feedback on gender differences in competitive choices. *Available at SSRN 1976073*.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–61.

APPENDICES

A Single-Period Model Extensions and Formal Results

This appendix collects full versions of propositions and formal statements referenced in the main theory section. Where relevant, we restate results for clarity and provide brief proof sketches.

A.1 Model Assumptions

(A1) Outputs y are i.i.d. across agents from distribution F.

(A2) Agents observe their own output but not others'.

(A3) Belief distributions $I(r)$ are discrete and defined over ranks $r \in \{1, \dots, N\}$.

(A4) Agents are risk-neutral with respect to payoffs.

(A5) The principal is passive and does not set incentives.

A.2 Choice of Earnings Function

The choice of the earnings (payoff) function is critical in experiments involving belief elicitation, as it directly shapes participant incentives and affects the interpretability and generalisability of results. Below, four common types of payoff structures are clearly outlined, each associated with distinct theoretical properties and real-world analogues.

1. Proper Scoring Rules (Accuracy-based). Proper scoring rules such as the Quadratic, Brier, or Logarithmic scoring rule reward accuracy by incentivising participants to truthfully report their subjective beliefs. The payoff explicitly depends on how close stated probabilities match observed outcomes (*see* e.g. Armantier and Treich (2013); Hossain and Okui (2013); and Schotter and Trevino (2014) for an overview). For example, the Quadratic Scoring Rule commonly used in experimental economics is:

$$\text{Payoff} = C - \sum_{r} (I(r) - \mathbb{1}\{r = r^*\})^2,$$

where C is a fixed constant, and $I(r)$ is the reported belief distribution over possible outcomes. Proper scoring rules are frequently applied in controlled experimental environments that aim to accurately elicit probabilistic beliefs.

2. Outcome-dependent Binary Payments (Rank-based Accuracy). A simpler and common standard alternative, used in for example Harrison et al. (2015), is a binary payoff conditioned explicitly on correctly identifying the exact rank or outcome:

$$\text{Payoff} = \begin{cases} y, & \text{if } \widehat{r} = r^*, \\ 0, & \text{otherwise.} \end{cases}$$

This type of payoff directly incentivises precise identification or prediction of outcomes, reflecting real-life settings where binary or discrete rewards depend solely on correct predictions, such as betting or promotional competitions.

3. Linear or Piecewise-linear incentive structures (Accuracy-proportional). Another standard alternative is a linear or piecewise-linear payoff structure based on accuracy, where participants receive payoffs proportional to the closeness of their guess to the actual outcome (e.g. Offerman et al., 2009). An example might involve paying participants based on how close their guess $\widehat{r}$ is to the actual rank r^*:

$$\text{Payoff} = y - \alpha \cdot |\widehat{r} - r^*|,$$

where α is a penalty parameter. This approach reflects real-world situations where rewards or penalties scale gradually with accuracy, such as performance bonuses or penalty systems.

4. Proportional Confidence-based Payments (Non-standard, Confidence-focused). The current experimental setup employs a non-standard proportional payoff structure:

$$\text{Payoff} = y \cdot I(r^*),$$

rewarding participants proportionally to their stated confidence in their true rank. This design captures scenarios in which rewards in real life depend on individuals' subjective confidence, influencing their allocation of effort, task devotion, or commitment. While this deviates from traditional accuracy-based schemes, it specifically accommodates psychological and behavioural mechanisms related to self-confidence and belief management.

To the author's knowledge, this payoff structure is novel and specifically crafted for the present study's psychological and behavioural emphasis on self-confidence and belief management.

Each payoff type thus suits different experimental and real-world contexts, and the proportional confidence-based payoff chosen here highlights the psychological dimensions central to this research.

A.3 Optimality Existence

Lemma A.1 (Existence of an Optimal Belief Allocation)**.** *In the one-shot environment, the agent's expected utility maximization problem admits a solution. That is, there exists an optimal belief distribution $\{I(r)\}_{r=1}^N$ and associated modal guess $\widehat{r}$ that maximizes $\mathbb{E}_{r^*\sim\pi}[U^W(r^*)]$.*

Proof Sketch. The agent's feasible set is the N-dimensional simplex $\Delta^N = \{I(r) \geq 0 \mid \sum_r I(r) = 1\}$, intersected with the condition that a unique $\widehat{r} = \arg\max_r I(r)$ exists. This set is compact and closed in finite-dimensional space. The agent's utility function $U^W(r^*)$ is continuous in $I(r)$ for each r, and thus the expected utility is continuous in the belief distribution. The maximization of a continuous function over a compact set guarantees the existence of a solution. □

A.4 Comparative Statics

Proposition (Restatement of Proposition 3.1)**.** *Suppose feedback is private. Then:*

- *Increasing α^+ (the cost of overestimation) weakly increases $\widehat{r}^*$: the agent becomes more conservative, avoiding high (overconfident) guesses.*

- *Increasing α^- (the cost of underestimation) weakly decreases $\widehat{r}^*$: the agent becomes more assertive, avoiding low (underconfident) guesses.*

In a discrete rank setting, we can examine the agent's expected payoff for each possible guess $\widehat{r} \in \{1, \ldots, N\}$, substituting in the mismatch cost terms and weighting by $\pi(r)$. If α^+ grows, the penalty from overconfidence ($r^* > \widehat{r}$) intensifies, so the best $\widehat{r}^*$ either remains the same or shifts upward to reduce $\Pr(r^* > \widehat{r})$. An analogous argument applies to α^-, γ^+, and γ^-.

Proof. Let the agent's expected utility incorporate asymmetric ego costs: overestimation ($r^* > \widehat{r}$) penalized by α^+, underestimation ($r^* < \widehat{r}$) penalized by α^-. Increasing α^+ raises the marginal cost of guessing too high, so the utility-maximizing $\widehat{r}^*$ either remains the same or shifts downward. Analogously, increasing α^- makes low guesses more costly, so $\widehat{r}^*$ shifts upward or stays the same. A piecewise derivative or subgradient approach in the continuous approximation yields the same conclusion. □

Proposition (Restatement of Proposition 3.2)**.** *Suppose the agent's guess $\widehat{r}$ is publicly observed. Then:*

- *Increasing γ^+ (the cost of appearing overconfident) weakly increases $\widehat{r}^*$.*
- *Increasing γ^- (the cost of appearing underconfident) weakly decreases $\widehat{r}^*$.*

Proof. The logic mirrors that of the private feedback case. Public mismatch costs shift perceived audience reactions rather than self-perception. Raising γ^+ makes appearing overconfident more costly, encouraging the agent to lower their guess. Raising γ^- has the opposite effect. Since payoffs are linear in mismatch penalties and the agent chooses $\widehat{r}^*$ optimally, these comparative statics follow directly from directional changes in expected utility. □

A.5 Gender Heterogeneity in the One-Shot Model

This subsection presents the full comparative statics and heterogeneity results referenced in Section 3.5. We allow mismatch sensitivities $(\alpha_g^+, \alpha_g^-, \gamma_g^+, \gamma_g^-)$ and prior beliefs $\pi_g^0(r)$ to vary by group $g \in \{m, f\}$, and derive the resulting implications for optimal guesses $\widehat{r}^*$.

Differences in Mismatch Sensitivities

Proposition A.2 (Gendered reactions in $\widehat{r}^*$ to α^+ and α^-)**.** *Suppose men and women face identical belief distributions $\pi(r)$ but differ in mismatch parameters (α_g^+, α_g^-). If $\alpha_m^+ > \alpha_f^+$ and $\gamma_m^+ = \gamma_f^+$ (all else equal), then men's optimal guess $\widehat{r}_m^*$ exceeds women's $\widehat{r}_f^*$ in any environment where private feedback is anticipated.*

Proof sketch. Follows directly from Proposition 3.1. An increase in α^+ raises the cost of overestimation; whichever group faces higher α^+ chooses a strictly higher rank guess *ceteris paribus.* □

Proposition A.3 (Gendered Reactions in $\widehat{r}^*$ to γ^+ and γ^-)**.** *Suppose men and women face identical belief distributions $\pi(r)$ but differ in mismatch parameters (γ_g^+, γ_g^-). If $\gamma_f^+ > \gamma_m^+$ with $\alpha_f^+ = \alpha_m^+$ (all else equal), then women choose a higher (worse) optimal guess such that $\widehat{r}_f^* > \widehat{r}_m^*$ whenever public feedback is anticipated.*

Proof sketch. Follows directly from Proposition 3.2. A higher γ^+ increases the cost of appearing overconfident. The group with the larger γ^+ will avoid low (better) rank guesses to reduce the risk of visible mismatch. □

Differences in Prior Beliefs

Proposition A.4 (Gender Differences from Biased Priors)**.** *If men and women have identical mismatch parameters but differ in priors $\pi_g^0(r)$ such that men place systematically more mass on top ranks, then men will (weakly) select a better (lower) modal rank $\widehat{r}_m^*$ than women* ceteris paribus*.*

Proof sketch. If $\pi_m^0(r)$ stochastically dominates $\pi_f^0(r)$, then posterior beliefs $\pi_m(r)$ will place more weight on better ranks after observing performance. The agent maximizes expected utility by aligning belief mass and modal guess with high-probability ranks, implying that $\widehat{r}_m^*$ is lower than $\widehat{r}_f^*$. □

Each result follows directly from the structure of the one-shot model in Section 3.3. For ease of exposition, we refer to types m and f (men and women), but the framework is general to any heterogeneity in beliefs or mismatch parameters.

B Repeated-Period Model Extensions and Formal Results

This appendix complements the repeated-round model in Section 3.6 by providing formal propositions and proof sketches that characterize belief drift and learning dynamics under partial feedback integration.

B.1 Monotonic Drift under Repeated Mismatch

Lemma B.1 (Monotonic Adjustment under Repeated Overestimation)**.** *Suppose the agent faces private feedback and in each round t she discovers that $r_t^* > \widehat{r}_t$ (i.e. she was overplacing). If $\eta^+ > 0$, then her sequence of guesses $\{\widehat{r}_t\}_{t=1}^T$ is* weakly increasing*, meaning $\widehat{r}_{t+1} \geq \widehat{r}_t$ for all t with strict inequality if the mismatch is strictly positive.*

Proof sketch. By the partial-update rule (Section 3.6), if round t ends with $r_t^* > \widehat{r}_t$, the agent sets

$$\widehat{r}_{t+1} \;=\; \widehat{r}_t + \eta^+ \left(r_t^* - \widehat{r}_t\right),$$

as long as $\eta^+ > 0$ and feedback is received. Because $r_t^* - \widehat{r}_t > 0$, the increment $\eta^+(r_t^* - \widehat{r}_t)$ is nonnegative. Hence $\widehat{r}_{t+1} \geq \widehat{r}_t$. Repeating this logic for each round $t = 1, \ldots, T-1$ yields a weakly increasing sequence. If the mismatch $(r_t^* - \widehat{r}_t)$, is strictly positive and $\eta^+ > 0$, the inequality is strict. □

Interpretation: As long as the agent internalizes some fraction $\eta^+ > 0$ of the "bad news," her belief report will drift upward across rounds when she consistently overplaces.

Lemma B.2 (Monotonic Adjustment under Repeated Underestimation)**.** *Suppose the agent faces private feedback and in each round t she discovers that $r_t^* < \widehat{r}_t$* (i.e., she was underestimating)*. If $\eta^- > 0$, then her sequence of guesses $\{\widehat{r}_t\}_{t=1}^T$ is* weakly decreasing*, i.e., $\widehat{r}_{t+1} \leq \widehat{r}_t$ for all t, with strict inequality if the mismatch is strictly negative.*

Proof sketch. The argument mirrors Lemma B.1. From the partial-update rule, when $r_t^* < \widehat{r}_t$, the agent adjusts according to

$$\widehat{r}_{t+1} = \widehat{r}_t + \eta^- \cdot (r_t^* - \widehat{r}_t).$$

Since the term $(r_t^* - \widehat{r}_t)$ is negative and $\eta^- > 0$, each adjustment is nonpositive, implying $\widehat{r}_{t+1} \leq \widehat{r}_t$. Strict inequality arises when the mismatch is strictly negative and internalized to any degree. □

B.2 Asymmetric Belief Updating Across Mismatch Types

Proposition B.3 (Sensitivity to Positive vs. Negative Feedback)**.** *Let agents A and B experience the same sequence of rank mismatches $\{r_t^* - \widehat{r}_t\}$ over $t = 1, \ldots, T$, but differ only in their update parameters η_g^+ and η_g^-.*[65] *Then:*

1. *If $r_t^* > \widehat{r}_t$ in every round (overplacement), and $\eta_A^+ > \eta_B^+$, then $\widehat{r}_T^A > \widehat{r}_T^B$.*
2. *If $r_t^* < \widehat{r}_t$ in every round (underplacement), and $\eta_A^- > \eta_B^-$, then $\widehat{r}_T^A < \widehat{r}_T^B$.*

[65] Representing η_A^+, η_B^+ and η_A^-, η_B^- respectively.

Proof sketch. In the overplacement case, each agent updates according to:

$$\widehat{r}^{g}_{t+1} = \widehat{r}^{g}_{t} + \eta^{+}_{g}(r^{*}_{t} - \widehat{r}^{g}_{t}).$$

A higher η^{+}_{A} implies that agent A adjusts upward more strongly than B in each round, yielding $\widehat{r}^{A}_{T} > \widehat{r}^{B}_{T}$ by induction. The underplacement case follows analogously, with negative mismatch and parameters η^{-}_{g}. □

Interpretation. When mismatch feedback consistently points in one direction, agents who internalize more of that signal adjust their beliefs further. This structure allows for type-dependent updating—e.g., one group may respond more to positive signals (underplacement), while another responds more to negative signals (overplacement). These asymmetries can be mapped to observed gender differences in feedback responsiveness.

Corollary B.3.1 (Permitting Gender Differences in Asymmetric Feedback Responsiveness)**.** *Suppose $\eta^{+}_{f} > \eta^{+}_{m}$ and $\eta^{-}_{f} < \eta^{-}_{m}$. Then the model permits an asymmetric pattern, under consistent mismatch exposure, in which:*

- *Women adjust their beliefs $\widehat{r}_t$, more in response to overplacement (negative feedback),*
- *Men adjust their beliefs $\widehat{r}_t$, more in response to underplacement (positive feedback).*

Proof sketch. Proposition B.3 shows that agents with higher η^{+} or η^{-} adjust more strongly in response to consistent mismatch in that direction. Since η^{+} and η^{-} are distinct parameters, nothing in the model restricts one agent from having a higher η^{+} but a lower η^{-} than another. It is therefore possible for agent f to adjust more in response to overplacement, while agent m adjusts more in response to underplacement. The model thus accommodates directional asymmetry in feedback responsiveness across agents or groups. □

This theoretical structure motivates an empirical test for asymmetric feedback responsiveness, as described in Corollary B.3.1.

B.3 Differential Weight on Performance vs. Mismatch Signals

Proposition B.4 (Differential Weight on Ability vs. Mismatch Feedback)**.** *Let $\delta^{y} \in [0,1]$ represent the weight the agent places on changes in own performance y_t, and let $1-\delta^{y}$ reflect the weight placed on discovered mismatch $(r^{*}_{t} - \widehat{r}_{t})$.*[66] *Suppose agents A and B follow the same updating rule but differ in δ^{y} and $\eta^{\pm}$. Then even when exposed to the same performance and feedback sequence, their belief paths $\{\widehat{r}_t\}$ may diverge, depending on how they interpret and prioritize these signals.*

Proof sketch. Let the agent's belief update rule be:

$$\widehat{r}_{t+1} = \delta^{y} \cdot \Psi(\widehat{r}_{t}, y_{t+1} - y_{t}) \; + \; (1-\delta^{y}) \cdot \Omega(\widehat{r}_{t}, r^{*}_{t} - \widehat{r}_{t}),$$

[66] In empirical settings, δ^{y} may reflect an agent's trust in the objectivity or relevance of performance-based metrics, while $\eta^{\pm}$ captures the willingness to internalize comparative feedback.

where $\Psi(\cdot)$ captures the influence of performance changes and $\Omega(\cdot)$ reflects partial mismatch adjustment (driven by η^+ or η^- depending on direction). Agent A places more weight on Ψ (i.e., $\delta_A^y > \delta_B^y$), while agent B places more weight on mismatch correction and has higher η parameters.

Even when the same sequence $\{y_t, r_t^*\}$ is observed, the agents' belief paths diverge: A responds more to performance variation, B responds more to mismatch feedback. □

Interpretation. The agent balances two main signals in each round: changes in own performance $(y_{t+1} - y_t)$ and the mismatch between reported and true rank. The model allows agents to vary in how they prioritize these sources—yielding different learning dynamics even under identical objective feedback. If she places greater weight on performance, she may override mismatch feedback; if she trusts the rank signal more, she adjusts accordingly. This structure captures the idea that individuals may prioritize one source of information over the other—generalizing how some may discount "rank error" in favour of absolute performance.

Corollary B.4.1 (Permitting Gender Differences in Updating Priorities)**.** *Suppose agents differ in their relative weighting of performance signals vs. mismatch feedback. If men place more weight on changes in own performance ($\delta_m^y > \delta_f^y$), while women place more weight on mismatch feedback (i.e., $\eta_f^+ > \eta_m^+$ or $\eta_f^- > \eta_m^-$), then belief paths $\{\widehat{r}_t\}$ will diverge, even under identical observed performance and feedback histories.*

Proof Sketch. Follows from Proposition B.4. Given the same sequences $\{y_t, r_t^*\}$, an agent who places greater weight on performance changes (δ^y) and lower weight on feedback responsiveness ($\eta^{\pm}$) will adjust beliefs primarily in response to y_t. The converse holds when mismatch feedback is prioritized. Thus, even under identical signals, belief paths diverge if agents differ in how they weigh the two sources. □

Concluding Remark. The results above rely on the linear partial-update structure described in Section 3.6, where belief paths evolve through a weighted combination of mismatch feedback and performance signals. Lemmas B.1—B.2 and Propositions B.3–B.4 formalize how belief trajectories respond to directional mismatch and agent-specific updating weights. A fully rigorous dynamic program—where agents anticipate future mismatch and optimize over belief paths—could be developed using backward induction. However, our conceptual framework is sufficient to highlight how asymmetric or incomplete incorporation of feedback drives belief dynamics across rounds.

C Sampling Methods

C.1 Data exclusion

Table C.1: Round 1: final data—by gender and treatment conditions

	Control	Private	Public	Joint	Total
Male	69	70	71	71	281
excluded	(3)	(2)	(2)	(1)	(8)
Female	68	67	68	70	273
excluded	(4)	(4)	(3)	(2)	(13)
Total	137	137	139	141	554
excluded	(7)	(6)	(5)	(3)	(21)

Table C.2: Round 2: final data—by gender and treatment conditions

	Control	Private	Public	Joint	Total
Male	66	66	66	66	263
excluded	(6)	(6)	(7)	(7)	(26)
Female	69	68	69	69	275
excluded	(3)	(3)	(2)	(3)	(11)
Total	135	134	135	134	538
excluded	(9)	(9)	(9)	(10)	(37)

Table C.3: Round 3: final data—by gender and treatment conditions

	Control	Private	Public	Joint	Total
Male	67	63	68	65	263
excluded	(5)	(9)	(5)	(7)	(26)
Female	68	65	66	66	265
excluded	(4)	(6)	(5)	(6)	(21)
Total	135	128	134	131	528
excluded	(9)	(15)	(10)	(13)	(47)

D Summary Statistics

D.1 Summary Statistics by gender

Table D.1 shows demographic and academic characteristics surveyed in the experiment introduction, together with a measures of the subjects' experimental understanding and economic situation compared to peers. The average sample age is 24 years and 7 months. Men are on average slightly older (≈ 3 months); have higher completed study years (≈ 1 month); are more likely to major in a STEM field (diff. 4 p.p.); and are about five percentage points less likely to major in languages, linguistics or literature (denoted as LLL), compared to women, but none of the differences are large or statistically significant.[67] Women speak 0.27 languages more on average than do men (p-value< 0.001).[68] The average self-reported economic situation compared to peers, is 4.9, on a scale from 0, denoted as "worse" and 10 "better" for the full sample and the gender difference is small and in significant. The proxy measure for experimental understanding, shows that four out the of five comprehension questions were correctly answered on the first attempt.[69]

Table D.1: Descriptive statistics: demographics and comprehension, by gender

	Scale	(1) All	(2) Male	(3) Female	(4) p-value
Age	18-51	24.59	24.70	24.48	0.528
		(4.10)	(4.15)	(4.07)	
Higher educ. (full yrs.)	0-6	3.33	3.27	3.39	0.420
		(1.78)	(1.86)	(1.70)	
Study-field major					
Majors in STEM	0/1	0.14	0.16	0.12	0.200
LLL	0/1	0.11	0.09	0.14	0.106
No. spoken languages	2-7	3.14	3.01	3.28	0.001
		(0.92)	(0.89)	(0.94)	
Economic standing	0-10	4.91	5.00	4.82	0.276
		(2.00)	(2.00)	(2.00)	
Comprehension (proxy)					
Correct 1st attempt	0-5	4.02	4.08	3.95	0.185
		(1.19)	(1.19)	(1.20)	
N		575	289	286	575

NOTES: The table shows in column 1-3 covariate averages of the full sample (all treatment conditions) and for males and females separately. Column 4 reports p-values from tests of gender mean differences: t-tests for continuous variables and χ^2-tests for binary variables.

[67] *See* Table D.4 for a full list of the study field majors and the classifications of STEM and LLL.

[68] Averages consider all reported proficiency levels: "Native"; "Fluent"; "Very good"; "Good" or; "Basic".

[69] Subjects cannot proceed without answering all questions correctly: a 0 indicates that more attempts were needed for (5) questions, and a 5 that all questions were immediately correct, not using hints or retrials.

Table D.2: Summary statistics and balance test—by treatment conditions

	Joint F-test	Treatment Conditions				Pairwise t-test		
		Control	Private	Public	Joint			
	p(F)	mean/(sd)				p-value		
	[(1):(4)]	(1)	(2)	(3)	(4)	(1)-(2)	(1)-(3)	(1)-(4)
Age	0.171	24.48	24.48	25.22	24.19	0.993	0.159	0.523
		(3.90)	(3.54)	(4.96)	(3.84)			
Higher educ. (full yrs.)	0.781	3.31	3.23	3.45	3.32	0.709	0.524	0.965
		(1.85)	(1.85)	(1.73)	(1.69)			
Study-field major								
STEM	0.177	0.09	0.13	0.17	0.17	0.331	0.053	0.053
LLL	0.048	0.17	0.07	0.11	0.10	0.007	0.129	0.088
No. spoken languages	0.469	3.18	3.07	3.10	3.22	0.324	0.452	0.703
		(0.97)	(0.93)	(0.91)	(0.89)			
Economic situation	0.866	4.92	4.83	4.88	5.02	0.676	0.862	0.671
		(1.96)	(2.02)	(2.11)	(1.92)			
Comprehension (proxy)								
Correct 1st attempt	0.524	4.11	3.95	4.07	3.94	0.268	0.764	0.199
		(1.15)	(1.29)	(1.20)	(1.14)			
N	575	144	143	144	144	287	288	288

NOTES: The table shows in column 2–5 mean average summary statistics for all treatment conditions: Control, Private, Public, and Joint using the full sample ($N = 575$). Column 1 shows p-values, p(F) per variable from joint F-test of $Control = Private = Public = Joint$. Columns 6–8 report p-values from pairwise tests of covariate mean differences, comparing the *Control* condition with each treatment condition: t-tests for continuous variables and χ^2-tests for binary variables. Standard deviations are in in parentheses.

D.2 Summary Statistics and Balance Testing

Table D.3: Summary statistics and balance test, Round 1–3 (constraints)—by treatment

	Joint F-test	Treatment Conditions				Pairwise t-test		
		Control	Private	Public	Joint			
	p(F)	mean/(sd)				p-value		
	[(1):(4)]	(1)	(2)	(3)	(4)	(1)-(2)	(1)-(3)	(1)-(4)
Panel A								
Age	0.097	24.32	24.58	25.06	23.99	0.523	0.123	0.391
		(3.25)	(3.54)	(4.61)	(3.13)			
Higher educ. (full yrs.)	0.909	3.36	3.30	3.44	3.31	0.767	0.722	0.811
		(1.84)	(1.82)	(1.75)	(1.68)			
Study-field major								
STEM	0.195	0.09	0.13	0.16	0.17	0.247	0.075	0.040
LLL	0.102	0.17	0.07	0.12	0.11	0.016	0.210	0.137
No. spoken languages	0.433	3.15	3.08	3.06	3.23	0.563	0.458	0.462
		(0.94)	(0.94)	(0.88)	(0.90)			
Economic situation	0.794	4.89	4.80	4.86	5.04	0.715	0.911	0.535
		(1.96)	(2.01)	(2.06)	(1.93)			
Comprehension (proxy)								
Correct 1st attempt	0.693	4.11	3.97	4.09	3.99	0.350	0.910	0.385
		(1.16)	(1.29)	(1.19)	(1.07)			
N	554	137	137	139	141	274	276	278
Panel B								
Age	0.101	24.43	24.54	25.13	23.95	0.804	0.189	0.269
		(3.97)	(3.62)	(4.68)	(3.11)			
Higher educ. (full yrs.)	0.688	3.22	3.28	3.47	3.34	0.812	0.251	0.584
		(1.83)	(1.88)	(1.70)	(1.63)			
Study-field major								
STEM	0.241	0.10	0.13	0.17	0.17	0.331	0.074	0.070
LLL	0.043	0.18	0.07	0.11	0.11	0.006	0.120	0.126
No. spoken languages	0.404	3.17	3.08	3.09	3.25	0.437	0.457	0.489
		(0.91)	(0.95)	(0.88)	(0.89)			
Economic situation	0.883	4.96	4.85	4.88	5.03	0.644	0.740	0.776
		(1.93)	(2.05)	(2.09)	(1.93)			
Comprehension (proxy)								
Correct 1st attempt	0.399	4.15	3.96	4.07	3.93	0.200	0.599	0.110
		(1.11)	(1.25)	(1.20)	(1.09)			
N	538	135	134	135	134	269	270	269
Panel C								
Age	0.081	24.44	24.55	25.16	23.92	0.812	0.180	0.237
		(3.98)	(3.50)	(4.69)	(3.12)			
Higher educ. (full yrs.)	0.726	3.27	3.28	3.49	3.33	0.949	0.312	0.765
		(1.85)	(1.87)	(1.74)	(1.63)			
Study-field major								
STEM	0.312	0.10	0.14	0.16	0.17	0.267	0.099	0.085
LLL	0.092	0.17	0.07	0.11	0.11	0.013	0.170	0.194
No. spoken languages	0.432	3.19	3.09	3.07	3.23	0.365	0.292	0.749
		(0.95)	(0.96)	(0.88)	(0.90)			
Economic situation	0.631	4.92	4.78	4.97	5.11	0.584	0.837	0.429
		(1.99)	(2.07)	(2.12)	(1.89)			
Comprehension (proxy)								
Correct 1st attempt	0.379	4.16	3.93	4.03	3.96	0.116	0.354	0.139
		(1.13)	(1.26)	(1.21)	(1.08)			
N	528	135	128	134	131	263	269	266

NOTES: The table shows in column 2–5 mean average summary statistics for all treatment conditions, using the full sample in each round (N = 554 for Round 1, N = 554 for Round 2, and N = 528 for Round 3). Column 1 shows p-values, p(F) per variable from joint F-test of $Control = Private = Public = Joint$. Columns 6–8 report p-values from pairwise tests of covariate mean differences, comparing the *Control* condition with each treatment condition: t-tests for continuous variables and χ^2-tests for binary variables. Standard deviations are in parentheses.

D.3 Study field classifications

Table D.4: Classifications of study fields

Major study field	**Men**	**Women**	**All**	STEM	LLL	Comment
Humanities and Social Sciences	148	153	301			
Economics	61	26	87			
Linguistics and languages	16	32	48	0	1	
Political science	12	33	45			
The arts	15	19	34			
History	15	6	21			
Philosophy	11	6	17			
Anthropology/Archaeology	8	7	15			
Interdisciplinary studies	0	11	11			
Literature	4	6	10	0	1	Recorded from "Other"
Psychology	3	3	6			
Sociology	2	3	5			
Geography	1	1	2			Recorded from "Other"
Religion	-	-	-			
Professions and Applied Sciences	91	87	178			
Engineering and technology	38	13	51	1	0	
Law	15	36	51			
Medicine	16	17	33			
Agriculture	10	1	11			
Education	3	8	11			
Journalism/media studies/communication	4	6	10			
Environmental studies and forestry	2	2	4			
Architecture and design	3	0	3			
Business	0	3	3			
Social work	0	1	1			
Library and museum studies	-	-	-			
Military sciences	-	-	-			
Public administration/Public policy	-	-	-			
Transportation	-	-	-			
Natural Sciences	21	24	45			
Biology	6	10	16	1	0	
Chemistry	5	7	12	1	0	
Physics	5	2	7	1	0	
Earth science	3	2	5	1	0	
Space science/Astronomy	2	3	5	1	0	
Formal Sciences	21	15	36			
Mathematics/Statistics	15	13	28	1	0	
Computer science	6	2	8	1	0	
Logic	-	-	-			
System science	-	-	-			
Other	8	8	16			
Biotechnology	0	1	1	1	0	Recorded from "Other"
Geology	1	0	1	1	0	Recorded from "Other"
other	7	7	14	0	0	
Total	289	287	576	(10)	(3)	

E Empirical Strategy and Variables

E.1 H11: Definition of Proportional Adjustment Variable

To analyse belief updating in response to feedback, I construct a proportional adjustment variable, `adjust_r2`, which captures how much of the previous round's feedback signal is incorporated into the agent's revised belief. Let $\widehat{r}_1$ and $\widehat{r}_2$ denote the agent's modal rank guess in round 1 and round 2, respectively, and let r_1 be the true realised rank in round 1. Then, the round-1 feedback signal is defined as:

$$\text{signal}_1 = \widehat{r}_1 - r_1,$$

where a negative value indicates overplacement and a positive value indicates underplacement.

The amount by which the agent updates their belief is $\widehat{r}_1 - \widehat{r}_2$, and the proportional adjustment is:

$$\text{adjust\_r2} = \frac{\widehat{r}_1 - \widehat{r}_2}{\widehat{r}_1 - r_1} = \frac{\text{belief change}}{\text{feedback signal}}.$$

This measure equals 1 when the agent fully corrects their prior error, 0 when no adjustment is made, and exceeds 1 in cases of overcorrection.

To handle cases where the signal is zero (i.e., the agent's initial belief was accurate), the following conventions are applied:

- `adjust_r2` is set to 1 if the agent does not adjust their belief after receiving accurate feedback.
- `adjust_r2` is set to -1 if the agent adjusts despite having been accurate.

This variable enables symmetric analysis of belief responsiveness conditional on the direction of feedback, and is used in Section 5.3 to test Hypothesis H11 on gender-specific updating behaviour.

F Pre-registry and model generated hypotheses

Table F.1: Mapping of testable hypotheses across main text, pre-analysis plan (PAP), and conceptual model.

Main ID	Label	PAP ID	Model Pred.	Description	Confirm.	Dir.	Comment
H1	Estimation Bias	H2A		Women overestimate RET scores less than men.	Yes	Yes	Based on pre-registration only.
H2	Belief Precision	H2C		Women allocate ECUs more broadly than men.	Yes	Yes	Related to uncertainty, not modelled.
H3	Placement Bias	H2B	PA	Women report worse (higher) modal ranks than men.	Yes	Yes	Matches both model and PAP.
H4	Actual Accuracy	H3		Women are equally likely to allocate to the true rank.	Yes	Yes	Measures extensive accuracy.
H5	Social-Image Effect	H1A	P1	Public exposure leads to more conservative self-assessment.	Yes	Yes	From model; also preregistered.
H6	Gender—Social Sensitivity	H1AG	P3	Women react more strongly to public observability.	Yes	Yes	Gender heterogeneity in model + PAP.
H7	Self-Image Effect	H1B	P2	Private feedback affects belief reporting.	Yes	No	From model; also preregistered.
H8	Gender—Self Sensitivity	H1BG	P4	Gender moderates self-image response.	Yes	No	No directional sign preregistered.
H9	Belief Updating	H4	P5	Belief adjusts in direction of feedback.	Yes	Yes	Part of feedback response analysis.
H10	Asymmetric Updating		P6	Stronger response to overplacement than underplacement.	Yes	Yes	Asymmetry in learning rates.
H11	Gender – Feedback Response	H4	P7	Women internalise negative feedback more.	Yes	Yes	From model and PAP.
H12	Exposure Interaction		P8	Public feedback may suppress learning.	Yes	Yes	Joint treatment effect.
–	Precision Level (exploratory)	H2D	–	Explore over-/under-precision across gender.	No	No	Not tested in main specification.

NOTES: Hypotheses in the main text (H1–H12) are mapped to corresponding identifiers from the preregistration (PAP ID) and the conceptual model (Model Prediction). The Confirm.-column indicates that the hypothesis has confirmatory status; and Dir. indicates whether a directional effect is hypothesised.

G Robustness checks and Other Specifications

G.1 Extended Descriptive Statistics for Round 1

Table G.1 presents the full treatment-level and gender-specific descriptive statistics underlying the Round-1 results discussed in Section 6.1.1. Real-effort performance is broadly similar across treatments. Agents scored on average 17.91 points in the *Control* condition (col. (1)), 17.54 in *Private* (col. (5)), 17.21 in *Public* (col. (9)), and 18.04 in *Joint* (col. (13)). Gender differences within treatments are small: for example, in *Control*, men scored 17.94 and women 17.88 (cols. (2)–(3), difference $p = 0.918$), and similar patterns appear in *Private* (17.64 vs. 17.44, cols. (6)–(7), $p = 0.765$), *Public* (17.68 vs. 16.75, cols. (10)–(11), $p = 0.123$), and *Joint* (18.65 vs. 17.46, cols. (14)–(15), $p = 0.061$). These figures confirm that performance is broadly balanced across gender and treatment cells in Round 1.

Score-belief estimation is based on an incentivised guess of the number of correctly solved items. Because agents know the number of attempted items, score overestimation is mechanically impossible whenever actual performance equals the number of attempts, generating upper truncation for a non-trivial share of participants. This design feature helps explain variation in overestimation rates across treatments in Table G.1 and motivates the robustness checks reported in Appendix G (see Tables G.2 and G.3).

Rank-allocation behaviour also varies across treatments and genders. On average, agents selected around five ranks in *Control* (5.07, col. (1)), *Private* (4.93, col. (5)), Public (5.29, col. (9)), and *Joint* (4.56, col. (13)). Only a small share selected a single rank, with values between 2 and 6% depending on treatment ("Only one" rows in Table G.1). Preferred (modal) ranks follow an intuitive treatment gradient: beliefs are more conservative in *Private* (7.88, col. (5)) and *Joint* (7.75, col. (13)) than in *Control* (5.50, col. (1)) and *Public* (6.64, col. (9)). Gender differences are also visible in the table: for instance, in *Control*, women choose a higher preferred rank than men (6.22 vs. 4.74, cols. (3) vs. (2)), and similar patterns appear in other treatments.

Placement accuracy—as measured by the difference between true and preferred rank—also differs across treatments. Underplacement is more frequent in *Private* (0.37, col. (5)) and *Public* (0.33, col. (9)), whereas overplacement is more common in *Control* (0.70, col. (1)) and *Public* (0.59, col. (9)). Accurate placement is relatively rare but more frequent in *Private* (0.16, col. (5)) and *Joint* (0.18, col. (13)) than in *Control* (0.05, col. (1)). These values provide the full numerical basis for the summary patterns reported in Section 6.1.1.

Table G.1: Descriptive Statistics Performance and Rank-Beliefs Allocation, by treatment and gender

		Control				Private				Public				Joint				Pairwise comp.		
	Scale	All (1)	Male (2)	Female (3)	p-val. (4)	All (5)	Male (6)	Female (7)	p-val. (8)	All (9)	Male (10)	Female (11)	p-val. (12)	All (13)	Male (14)	Female (15)	p-val. (16)	p-val. (1)-(5)	p-val. (1)-(9)	p-val. (1)-(13)
Real-effort task performance																				
Score	5–32	17.01	17.03	17.00	0.957	16.53	16.51	16.54	0.970	16.46	16.77	16.13	0.299	16.89	17.32	16.44	0.154	0.227	0.174	0.753
		(3.10)	(3.45)	(2.72)		(3.57)	(4.18)	(2.84)		(3.63)	(4.01)	(3.18)		(3.66)	(4.05)	(3.19)				
Score belief	2–31	17.11	17.13	17.09	0.938	16.50	16.51	16.49	0.973	16.18	16.75	15.59	0.080	16.99	17.45	16.51	0.127	0.150	0.031	0.763
		(3.17)	(3.42)	(2.93)		(3.75)	(4.23)	(3.21)		(3.89)	(4.16)	(3.53)		(3.64)	(3.99)	(3.20)				
Estimation	-12–4	0.09	0.10	0.09	0.929	-0.02	0.00	-0.04	0.831	-0.28	-0.03	-0.54	0.052	0.10	0.13	0.07	0.759	0.362	0.014	0.970
		(0.87)	(0.77)	(0.96)		(1.22)	(1.29)	(1.16)		(1.57)	(0.89)	(2.02)		(1.06)	(1.22)	(0.89)				
Underest.	0/1	0.14	0.13	0.15	0.778	0.21	0.20	0.22	0.732	0.22	0.15	0.29	0.049	0.20	0.15	0.24	0.191	0.112	0.069	0.183
Accurate est.	0/1	0.65	0.65	0.65	0.950	0.53	0.53	0.52	0.942	0.62	0.65	0.59	0.469	0.53	0.56	0.50	0.451	0.037	0.594	0.046
Overest.	0/1	0.21	0.22	0.21	0.869	0.26	0.27	0.25	0.814	0.16	0.20	0.12	0.199	0.27	0.28	0.26	0.743	0.320	0.253	0.260
Rank	1–18	8.27	8.16	8.38	0.787	9.08	9.11	9.04	0.940	8.65	8.32	8.99	0.448	8.43	7.76	9.11	0.130	0.190	0.528	0.789
		(4.81)	(5.06)	(4.57)		(5.39)	(6.09)	(4.58)		(5.11)	(5.26)	(4.97)		(5.30)	(5.42)	(5.12)				
Ranked first	0/1	0.07	0.07	0.07	0.981	0.08	0.09	0.07	0.811	0.08	0.08	0.07	0.811	0.06	0.08	0.04	0.312	0.820	0.847	0.762
Ranked last	0/1	0.04	0.04	0.03	0.661	0.07	0.14	0.00	0.001	0.06	0.06	0.07	0.681	0.03	0.06	0.00	0.044	0.184	0.285	0.702
Rank allocations																				
No. of ranks	1–18	5.15	5.06	5.25	0.708	5.49	5.71	5.25	0.424	5.15	5.39	4.90	0.315	5.40	5.31	5.49	0.718	0.382	0.995	0.488
		(2.99)	(1.94)	(3.78)		(3.35)	(3.48)	(3.23)		(2.91)	(2.96)	(2.85)		(2.87)	(2.78)	(2.99)				
Only one	0/1	0.04	0.01	0.07	0.091	0.03	0.00	0.06	0.038	0.04	0.03	0.06	0.374	0.01	0.00	0.01	0.312	0.519	0.980	0.051
Preferred rank	1–18	5.42	4.72	6.12	0.034	5.72	4.97	6.49	0.025	6.94	6.28	7.62	0.082	6.18	5.46	6.90	0.024	0.528	0.003	0.098
		(3.86)	(3.28)	(4.27)		(3.99)	(3.40)	(4.41)		(4.52)	(4.81)	(4.13)		(3.78)	(3.65)	(3.80)				
Share allocated[1]	0–1	0.37	0.34	0.40	0.044	0.36	0.33	0.39	0.060	0.37	0.36	0.39	0.275	0.35	0.35	0.35	0.906	0.552	0.974	0.330
		(0.19)	(0.13)	(0.23)		(0.17)	(0.13)	(0.20)		(0.18)	(0.17)	(0.18)		(0.14)	(0.12)	(0.15)				
Placement[2]	-16–16	2.85	3.43	2.26	0.235	3.36	4.14	2.55	0.103	1.71	2.04	1.37	0.535	2.26	2.30	2.21	0.929	0.461	0.119	0.371
		(5.75)	(5.08)	(6.33)		(5.71)	(5.99)	(5.33)		(6.37)	(6.12)	(6.65)		(5.39)	(5.21)	(5.61)				
Underplac.	0/1	0.26	0.22	0.31	0.224	0.26	0.23	0.28	0.461	0.31	0.27	0.35	0.277	0.35	0.34	0.37	0.679	0.890	0.392	0.098
Accurate plac.	0/1	0.04	0.04	0.04	0.985	0.09	0.06	0.13	0.123	0.09	0.11	0.06	0.258	0.06	0.06	0.06	0.984	0.096	0.152	0.622
Overplac.	0/1	0.69	0.74	0.65	0.243	0.65	0.71	0.58	0.105	0.60	0.62	0.59	0.704	0.59	0.61	0.57	0.680	0.440	0.121	0.069
N		137	69	68	137	137	70	67	137	139	71	68	139	141	71	70	141	274	276	278

[1]Agents are required to allocate at least 2 ECUs to their preferred rank, implying a minimum possible share of approximately 0.105 (from 2/19).

[2]Placement is defined as the *Actual Rank* minus the *Preferred Rank* (the modal belief). A positive value indicates overplacement (believing oneself to have performed better than in reality), a negative value indicates underplacement, and zero indicates an accurate rank belief.

Notes: This table reports round 1 descriptive statistics for agents' performance, belief accuracy, and rank allocation decisions. The upper panel covers RET scores and estimation beliefs, including indicators for under-, accurate-, and overestimation. The lower panel summarises rank selection behaviour, including the number of ranks selected, the preferred rank (modal belief), belief precision, and placement accuracy. Columns are grouped by treatment condition: *Control* (1–4), *Private* (5–8), *Public* (9–12), and *Joint* (13–16). Within each group, means are shown for the full sample (left), males (centre-left), and females (centre-right), with the rightmost column reporting p-values for gender differences using t-tests for continuous variables and χ^2-tests for binary outcomes. The final three columns report p-values from pairwise tests of mean differences between the *Control* and each treatment group (pooled across gender).

Pooled data

G.2 Hypothesis 1—Gender differences in (over)estimation

This appendix provides the full set of analyses underlying Hypothesis H1, which posits that women are less prone than men to overestimate their real-effort performance. Estimation accuracy is measured as the difference between an agent's incentivised score guess and her realised score in the round. Because overestimation is mechanically impossible when the actual score equals the number of attempted items, estimation errors are truncated from above for a subset of observations; this feature should be borne in mind when interpreting the magnitude of gender gaps.

In the full sample pooled across all rounds, women estimate their scores 0.16 points lower than men on average ($p < 0.01$) and are significantly less likely to overestimate their performance. A logistic model indicates that women have 19% lower odds of overestimating than men (odds ratio=0.81, $p = 0.037$; see Table G.2). These estimates likely understate the true gender difference because many agents were not at risk of overestimating in a given round.

Table G.2: Placement estimation: Performance Score - Pooled rounds

	Rounds 1-3 (pooled)		Individual avg.
	(1) Mean/(se.)	(2) Odds ratio[1]	(3) Mean/(se.)
Female	-0.155**	0.809*	-0.324*
	(0.054)	(0.082)	(0.140)
Constant	0.055	0.070***	0.573***
	(0.054)	(0.010)	(0.077)
Observations	1719	1719	328
Female (obs.)	857	857	157
Male (obs.)	862	862	171
Round controls	✓	✓	-
Est. types controls	-	✓	-

NOTES: The table shows coefficients from OLS regressions in column 1 and 3 and the Odds ratio from a logistic regression in column 2. Model (1) regress a binary gender dummy and round controls on *Estimation* (actual score - guessed score), using all three rounds. Model (2) regress a binary gender dummy and a variable counting number of rounds where all types of estimations were possible and round controls on a binary indicator for whether an agent overestimated its score performance. Both model 1 (and 2) uses the full sample and display means (odds ratio) and robust standard errors, clustered at individual level. Model (3) regress a gender dummy on the individual average estimation over the rounds in which all types of estimations were possible or if an agent had one round, only that value is used.

To address this censoring issue, I restrict the sample to rounds in which all three estimation outcomes—under-, accurate-, and overestimation—were feasible. The results are qualitatively identical: women report more conservative score guesses and remain significantly less likely than men to overestimate (Table G.3). Across both specifications, the findings strongly support H1, indicating systematically lower estimation confidence among women.

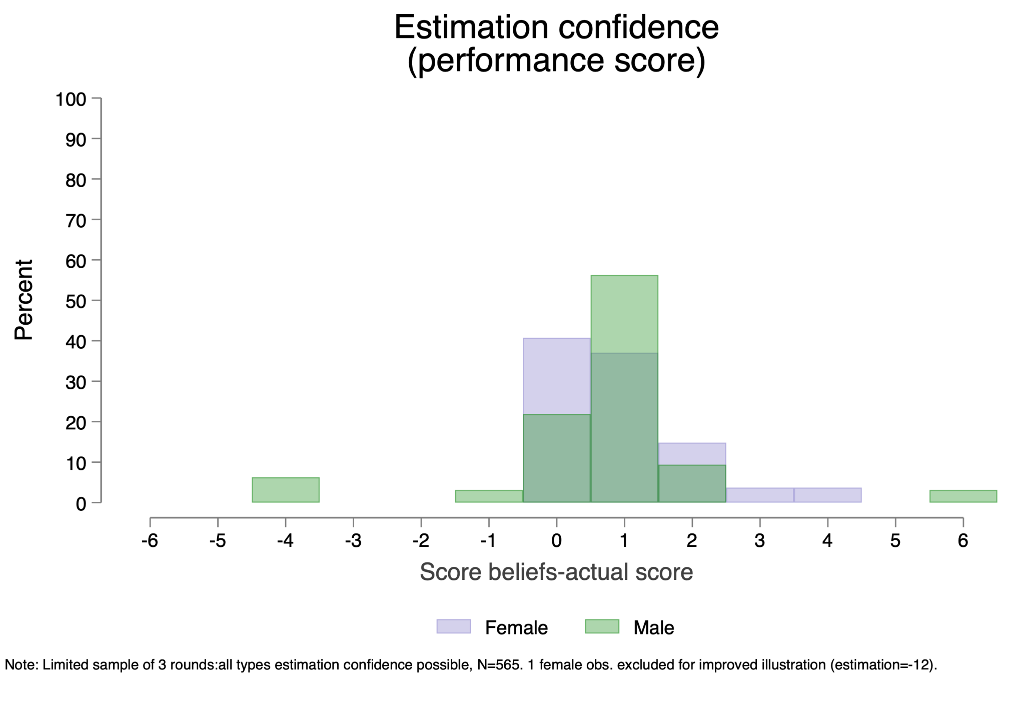
Table G.3: Placement estimation (restricted sample): Performance Score - pooled rounds

	Rounds 1-3 (pooled)		Individual avg.
	(1)	(2)	(3)
	Mean/(se.)	Odds ratio	Mean/(se.)
Female	-0.206*	0.573*	-0.324*
	(0.104)	(0.135)	(0.140)
Rounds - all est.[1]		0.430***	
		(0.070)	
Constant	0.924***	22.976***	0.573***
	(0.059)	(8.320)	(0.077)
Observations	529	529	328
Female (obs.)	252	252	157
Male (obs.)	277	277	171

[1]Number of experimental rounds (1–3), in which all the three types of estimation were possible, *i.e.* both over-, accurate, under-estimation.

NOTES: The table shows coefficients from OLS regressions in column 1 and 3 and the Odds ratio from a logistic regression in column 2. Model (1) regress a binary gender dummy and round controls on *Estimation* (actual score - guessed score), using all three rounds. Model (2) regress a binary gender dummy and a variable counting number of rounds where all types of estimations were possible and round controls on a binary indicator for whether an agent overestimated its score performance. Both Models (1) and (2) uses the full sample and display means (odds ratio) and robust standard errors, clustered at individual level. Model (3) regress a gender dummy on the individual average estimation over the rounds in which all types of estimations were possible or if an agent had one round, only that value is used.

Figure G.1: Performance estimation (restricted sample), all rounds



G.3 Hypothesis 2—Gender differences in (over)-precision

To test Hypothesis **H2**, I examine whether women display lower belief precision than men, measured by the number of ranks selected. Focusing on the *Control* group to avoid strategic adjustments due to anticipated exposure, I pool all three rounds and regress the number of selected ranks on gender, controlling for task performance and round fixed effects. Table G.4 shows that the gender difference is small and statistically insignificant. These results do not support **H2**.

Table G.4: Belief precision (control condition): pooled rounds

	Rounds 1–3 (pooled)	
	(1) Mean/(se.)	(2) Mean/(se.)
Female	0.206	0.207
	(0.380)	(0.382)
Constant	7.156***	7.087***
	(1.011)	(1.012)
Observations	407	407
Female (obs.)	205	205
Male (obs.)	205	205
RET score	✓	✓
Round controls		✓

NOTES: The table reports OLS estimates of gender differences in belief precision, measured by the number of ranks selected. The sample includes all agents in the Control condition across rounds 1–3. Column (1) controls for RET performance; column (2) additionally includes round fixed effects. Standard errors in parentheses, clustered at the agent level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

G.4 Hypothesis 3—Gender differences in (over)-placements

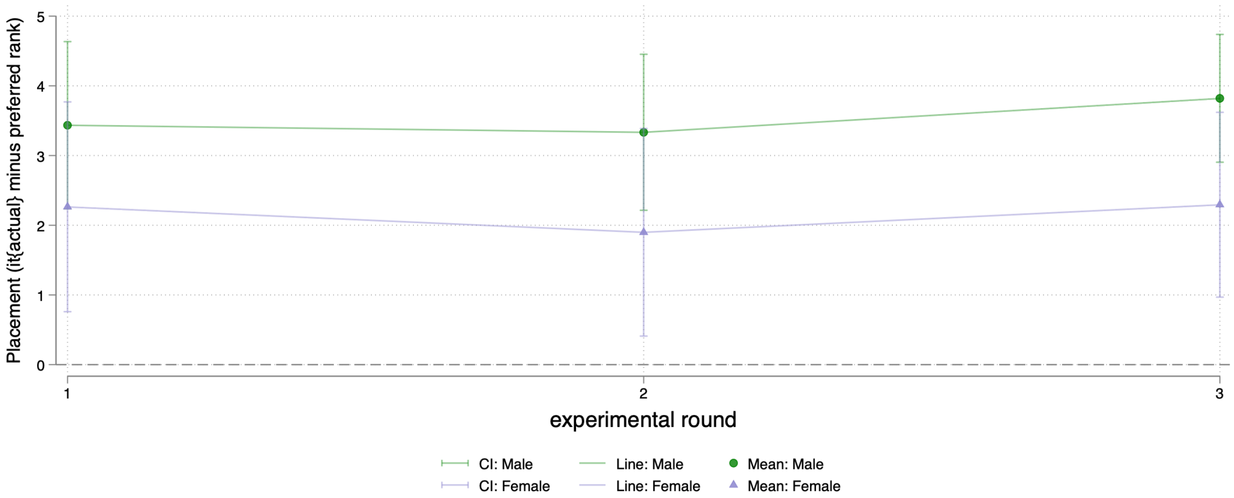
Table G.5: Self-confidence: placement (Control)—all rounds

	Rounds 1–3 (pooled)	
	(1) Mean/(se.)	(2) Mean/(se.)
Female	-1.492*	-1.506*
	(0.660)	(0.660)
Real-effort score	-0.932***	-0.995***
	(0.087)	(0.095)
Observations	407	407
Female (obs.)	205	205
Male (obs.)	205	205
RET score	✓	✓
Round controls		✓

NOTES: The table reports OLS estimates from regressions of agents' modal rank (preferred rank) on a gender indicator, using data from the Control condition across all three rounds. The dependent variable is the rank to which the agent allocated the most ECUs in each round. Higher values indicate worse (more conservative) self-assessments. Column (1) controls for RET performance (Real-effort score); column (2) additionally includes round fixed effects.

Gender differences in self-placement remain stable across rounds, with no indication that the gap narrows over time. As shown in Figure G.2, both men and women tend to overplace themselves on average in the *Control* condition, but the bias is consistently smaller for women—who report more conservative self-assessments throughout.

Figure G.2: Placement Bias by Gender and Round (Control Condition)



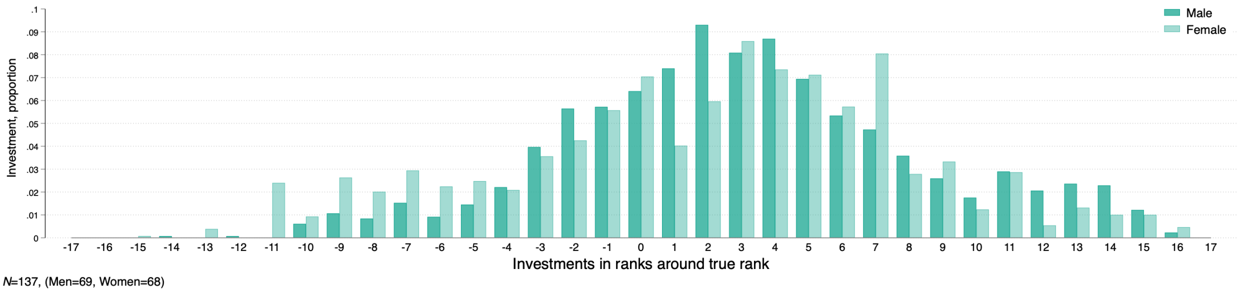
NOTES: The figure shows average placement bias, defined as the difference between true rank and the rank receiving the most ECU allocation, by gender and round. Positive values indicate underplacement (i.e., conservative self-assessment). Vertical lines denote 95% confidence intervals. Sample restricted to the Control condition across all three rounds.

H Additional tests of anticipation effects: Round 1

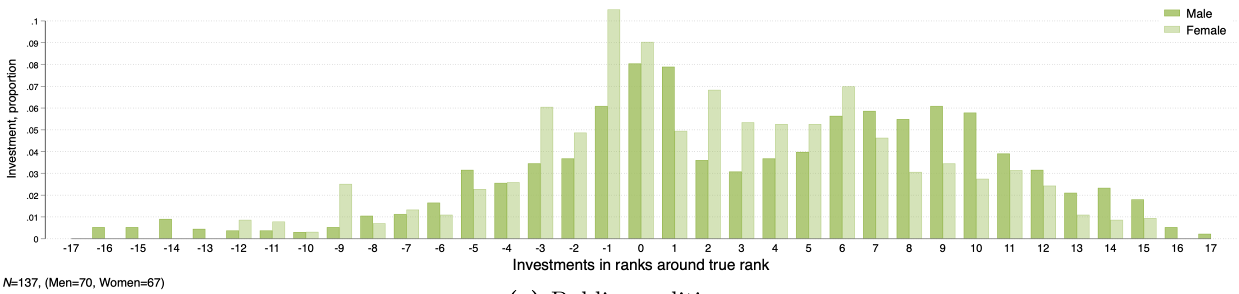
H.1 Rank allocations - all investments - round 1

Figure H.1: All Rank Investments, by Gender - first round

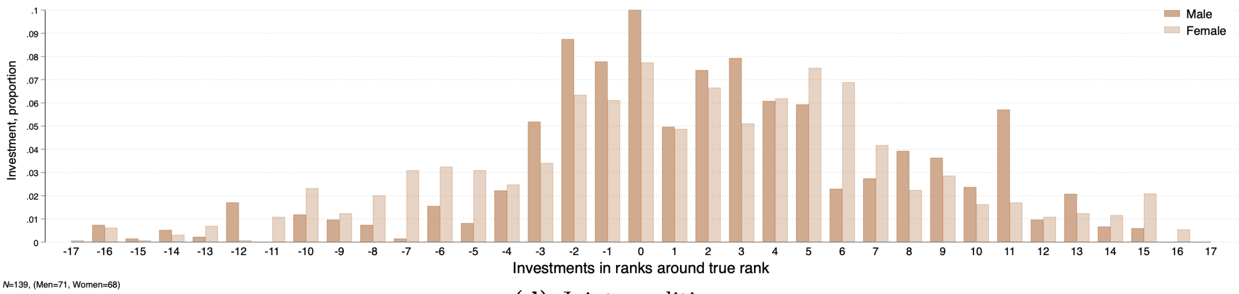
(a) Control condition



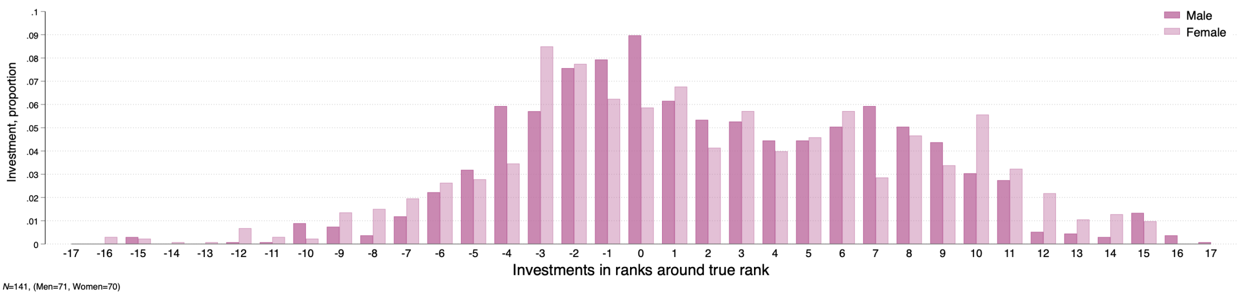
(b) Private condition



(c) Public condition

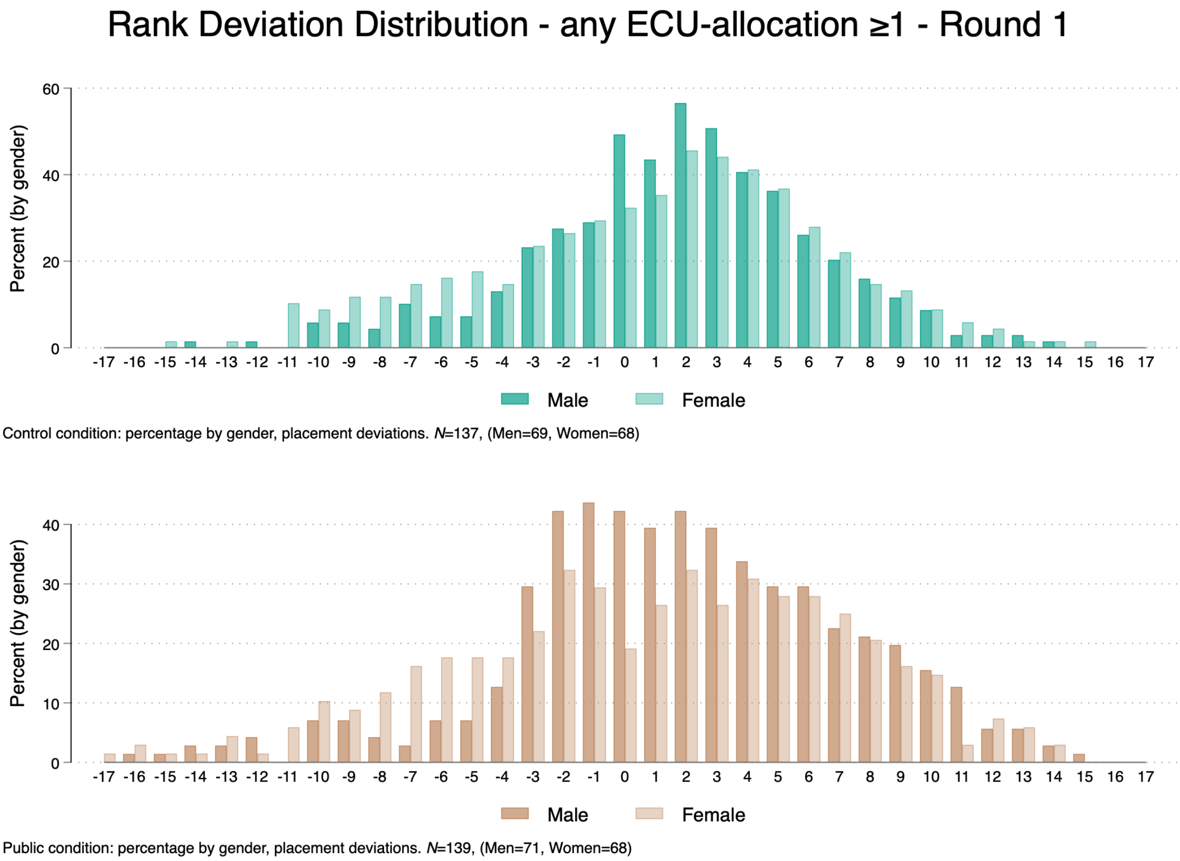


(d) Joint condition



NOTES: The figure shows additional ranks investments in rank-steps from the modal rank, as shares of the total remaining allotment - excluding the modal rank allocation (ECUs in the most preferred rank). The upper Figure 8a, displays investments for the *Control* condition, and the lower Figure 8b for the *Public* condition. Each treatment condition is separated by gender, showing females in lower intensity compared to males.

Figure H.2: Placement for all rank investment, by gender



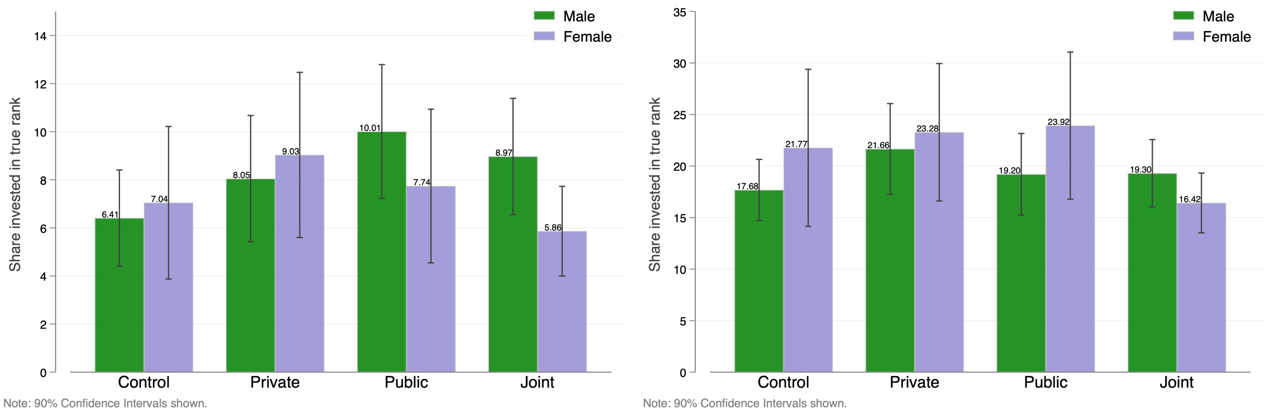
In the upper part of Figure H.2 the bars outlay a similar outcome of placements for women and for men and a t-test confirms that there is no statistically significant difference. The lower part of the figure confirms that only by making one of the rank-investments, the one that they invested the most in, comparable to their true rank, female agents shift down their belief distribution for their rank, considerably more than their male counterparts.

H.2 Extensive and intensive margin analysis of true-rank investment (Round 1)

This appendix provides supplementary analysis of how agents allocate belief weight to their true rank along both the extensive and intensive margins. These results complement the main-text findings in Section 6.1.4, which focus on the likelihood of selecting the true rank (extensive margin). Here, I document how much of the endowment is allocated to the true rank (intensive margin) and show that treatment effects mainly arise from the extensive margin rather than differences in belief intensity.

Figure H.3 illustrates how much of the endowment agents allocate to their true rank across treatments. Panel (a) shows the extensive margin, including all agents; Panel (b) shows the intensive margin, conditioning on agents who allocate at least one ECU to their true rank.

Figure H.3: Guessed true rank: all ranks selected, by gender - first round



(a) Extensive margin **(b)** Intensive margin

NOTE: The left Figure H.3a and the right Figure H.3b.

Along the extensive margin (Figure H.3a), men in the *Public* treatment allocate on average 3.6 percentage points more to their true rank than men in the *Control* group (10.01% vs. 6.41%, $p = 0.044$), corresponding to a 56% increase. No comparable differences are observed for women across any treatments. Along the intensive margin (Figure H.3b), allocations are statistically indistinguishable both within gender across treatments and within treatments across genders. This suggests that conditional on believing the true rank is possible, men and women allocate similar amounts to it.

In sum, the higher average allocations to the true rank among men in the *Public* treatment are primarily driven by a greater likelihood of assigning any belief weight to the true rank at all. Differences in belief intensity or certainty do not appear to drive these patterns.

I Additional tests of experiencing effects: round 2 or 3

Table I.1: OLS Robustness: Gender Differences in Feedback Adjustment (Round 2)

	Overplacement (signal < 0)	**Underplacement (signal > 0)**
Female	$0.422^{\dagger}$ (0.235)	-0.424^{**} (0.201)
Study years	0.057 (0.071)	0.010 (0.065)
STEM	–0.199 (0.258)	0.152 (0.157)
LLL	–0.354 (0.216)	–0.402 (0.516)
Constant	0.476^{*} (0.210)	0.896^{***} (0.289)
R^2	0.033	0.076
Observations	157	82

NOTES: Dependent variable is *adjust_r2*, defined as the proportion of the prior signal (Round 1), incorporated into the agent's updated stated modal rank belief in Round 2. Regressions are estimated separately for agents who received overplacement and underplacement feedback in round 1. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. All models include robust standard errors.

J Experimental Instructions

Figure J.1: page 1

Welcome to the experiment!

The experiment will soon begin!
From now on you will not be able to speak to the other participants or communicate in any other way. If you have **any questions** about the experiment, **please raise your hand** and we will attend to your working station as soon as possible.

Please **read all the instructions carefully before submitting any answers or leaving a page.**
During this experiment, no other aids are allowed apart from the scribble paper and the pen that are provided on your desk. Please mute your phone completely or turn it off and then put it away in a pocket or a bag that you cannot reach for the rest of the experiment.

We are very happy that you have **chosen to participate** and that you have filled the Consent Form. At the end of the experiment we will pay you a **show-up fee of 5 €** besides all other potential earnings of this experiment.
Please do not discuss the content of this experiment with anyone, inside or outside the lab!

Please click "Next" to continue reading about the experiment.

NEXT

Figure J.2: page 2

Experiment Structure

Earnings: During the experiment you will make different decisions, from which you can sometimes earn **"ECUs"** - the experiment currency. These ECU-earnings will be translated in to real money (euro) in the end of the experiment according to the following rule: **10 ECUs correspond to [1] €**

Parts: The experiment consists of three main parts, Part A, B and C. All parts will be explained in detail as you proceed. In some of the parts you will be grouped with other participants in the laboratory and you may have to wait for their answers. When this happens, there will be an indicating "waiting page".

Instructions: You can separately earn money from Part A and B in the experiment. You will receive further detailed instructions for all parts and if some of the parts are connected this will also be explained.

Click "Next" to proceed to the experiment.

NEXT

Figure J.3: page 3

Introductory questions

Select your age in years:

`Use drop-down list with year span:` *18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 50+*

Gender:
◯ Man ◯ Woman

Select your main field of study (select "none" if you never been a student; select "other" and specify if your main study area is not in the list):

`Drop-down list with educational tracks - List in two levels, where first level is not electable.`
`Humanities and Social Sciences: Anthropology/Archaeology; History; Linguistics and languages; Philosophy; Religion; The arts; Economics; Geography; Interdisciplinary studies; Political science; Psychology; Sociology`
`Natural Sciences: Biology; Chemistry; Earth science; Physics; Space science/Astronomy`
`Formal Sciences: Computer science; Logic; Mathematics/Statistics; System science`
`Professions and Applied Sciences: Agriculture; Architecture and design; Business; Education; Engineering and technology; Environmental studies and forestry; Journalism/media studies/ communication; Law; Library and museum studies; Medicine; Military sciences; Public administration/Public policy; Social work; Transportation`
`Other: Other; None`

Other:
Specify:

`Pop-up field if selected "other". If "none" selected, impossible to select any years of education.`

Number of finished years:
◯ 0 ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 5+

Please tick the languages that you speak apart from Italian and indicate your level of command:

◯ English	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Spanish	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ French	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ German	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Albanian	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Other	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Other	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic

`Use Fill-in for Other. Possible to tick more than one. Use 5 tick-boxes + 2 others with free text.`

Click "Next" to start **Part A** of the experiment.

NEXT

Figure J.4: page 4

Overview of Part A

Part A consists of two main tasks:

- A **"Decoding Task"**: solve problems for **4 minutes.**
- A **"Contract Selection":** choose a working contract that determines the pay you get for your correctly solved problems.

This is repeated in 3 rounds.

Part A has two roles: All participants will randomly be assigned one of the two **roles:** "Employee" or "Principal".

- **Employees:** perform 3 rounds of Decoding Tasks and Contract Selections.
- **Principals:** will be randomly **matched with 3 Employees**. The Principals will only perform the Decoding Task in Round 1.

Your assigned role shows up on your screen before the first Contract Selection.

The score, the rank and Selecting a Contract:

- **Score:** In the Decoding Task, each correctly solved problem gives one "point" and will be added up to the Employee's **score** of that round.
- **Rank:** All the 18 Employees' **scores** will automatically be **ranked** such that **rank 1** is assigned to the Employee with the **highest score** and so on until **rank 18** is assigned to the Employee with the **lowest score.** Note that in case there is a score tie between two or more Employees, everyone gets the same higher rank.
- **Contract Selection:** In the Contract Selection, **each Employee chooses** how to **distribute** a total of **[19] ECUs** among 18 available contracts named **Contract 1-18**.

Click "Next" to go to the overview of **How to earn money in Part A.**

Next

Figure J.5: page 5

Summary of "How to earn money" for Part A

- **Before the experiment started one of the three rounds has been randomly drawn for payment.** That round is announced at the very end of the experiment.
- **Any payment in Part A is separate from all other payments in the experiment.**

Earnings for Part A:

- **Employees:** You can earn ECUs only if you **choose a contract** that has the **same number** as your **rank**! The amount is then determined by the number of ECUs you chose to put on that contract, called the **"wage", times** your **score** in that round. All other contracts not having the same number as your rank, will give you 0 ECUs.

  Additionally, a secret **random pay** of 0, 1 or 2 times your score will be added to your final pay.

- **Principals:** earns ECUs from its Employees. A principal earns a **third** of **each** of the 3 Employees' earnings. This is the average pay of the Employees ,including the random pay. *Meanwhile the Employees select their contracts, the Principals will also be asked some questions from which they can earn more ECUs.*

Information in Part A

Once the Employees have done their contract selection the **Principals are provided information** on their **Employees'**:

- **actual rank** but neither the contracts they selected nor their score.
- In the end of the experiment Principals and Employees **learn their total earnings** of the round that was drawn for payment, but not what was added from the random pay.

General timeline of Part A

Round 1-3:

- Decoding Task;
- Role Information (only round 1);
- Contract Selection Task (only for Employees);
- *Answering Questions (only for Principals).*

More details, information, trial rounds and examples will be given in the instructions of each part.

Click "Next" to go to the instructions of the Decoding Task.

Next

(a) Control Treatment

Summary of "How to earn money" for Part A

- **Before the experiment started one of the three rounds has been randomly drawn for payment.** That round is announced at the very end of the experiment.
- **Any payment in Part A is separate from all other payments in the experiment.**

Earnings for Part A:

- **Employees:** You can earn ECUs only if you **choose a contract** that has the **same number** as your **rank**! The amount is then determined by the number of ECUs you chose to put on that contract, called the **"wage", times** your **score** in that round. All other contracts not having the same number as your rank, will give you 0 ECUs.
- **Principals:** earns ECUs from its Employees. A principal earns a **third** of **each** of the 3 Employees' earnings. This is the average pay of the Employees. Meanwhile the Employees select their contracts, the Principals will also be asked some questions from which they can earn more ECUs.

Information in Part A

Once the Employees have done their contract selection the **Principals are provided information** on their **Employees'**:

- **actual rank**, the **contracts** they selected but not their score.
- In the end of the experiment Principals and Employees **learn their total earnings** of the round that was drawn for payment.

General timeline of Part A

Round 1-3:

- Decoding Task;
- Role Information (only round 1);
- Contract Selection Task (only for Employees);
- *Answering Questions (only for Principals).*

More details, information, trial rounds and examples will be given in the instructions of each part.

Click "Next" to go to the instructions of the Decoding Task.

Next

(b) Public Treatment

Figure J.6: page 6

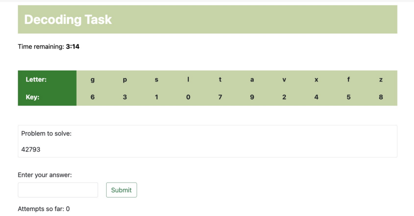
Decoding Task Instructions

The time period for a **Decoding Task** is **4 minutes**. Only a correct and submitted problem gives one point and incorrect answers do not give minus points. New problems appear automatically on the screen, as soon as an answer is submitted.

Example Screen

Below we show an example of a problem. You can see four things:

1) On the top row, the letters where you will search for your answer;
2) On the second row, the decoding key showing which number to match with which letter;
3) In the white box, the "Problem to solve" - the number series to translate to letters and;
4) The answer box where the letter combination should be entered before submitting the answer.



In the example, the problem to solve is to decode 42793. Following the decoding key, the 4 corresponds to the letter x, the 2 to the v, 7 to t, 9 to a and the 3 to the p. To get one point, xvtap needs to be entered in the answer-box and submitted.
At the top, a clock will indicate the remaining time and at the bottom, "Attempts so far" shows how many problems you have tried so far (both correct and incorrect).

Click "Next" to go to the **trial round** of the Decoding Task.

NEXT

Figure J.7: page 7

Trial round

Now you can **try** the Decoding Task for 2 minutes!

During this time period it does not matter if you are right or wrong: it will not affect any of your final earnings. In the trial round you can also see if your last answer was right or wrong, which will not be the case in the real rounds.

As soon as you click "Start Trial" the 2 minutes will start.

START TRIAL

Figure J.8: page 8

[TRIAL ROUND 2 min]

Figure J.9: page 9

Decoding Task

You have now finished the trial round. As soon as you click "Start Task" the 4 minutes of Decoding Task will start.

START TASK

Figure J.10: page 10

[DECODING TASK 4 min]

Figure J.11: page 11

Decoding Task summary

Congratulations, you have finished the [first/second/third] **Decoding Task** and you attempted **[X] problem[s]!**

Below we ask you [a] **question** about the **Decoding Task**. A correct answer is **rewarded with [10] ECUs**. As soon as you click "Submit" you will proceed to the next page.
How many of your [X] attempt[s] do you think were correct?

Enter your answer:

SUBMIT

`Cannot be larger than number of attempts and not smaller than 0.`

Figure J.12: page 12

Decoding Task

[Y] of your [X] attempt[s] were correct.

Click "Next" to go to the Contract Selection.

NEXT

Figure J.13: page 13

Contract Selection Instructions

Below we will explain how to select working contract(s) that will determine your earnings. All the 18 Employees repeat this task in all the three rounds.

Score rank: In each round, the Decoding Task scores will automatically be ranked, but are not revealed. The Employee with the **highest score** receives **rank 1** while the second highest scoring Employee receives **rank 2** etc. until the lowest scoring Employee may get **rank 18**. For example, a rank 6 means that 5 other Employees had a higher score. Note that the higher the score, the lower the rank!

Score ties: If two (or more) Employees have the **same score** they receive the **same rank** since there are as many Employees with **higher scores** compared to them. In the example table below you can also see that if there is a tie among two Employees with the very lowest score, the highest received rank is 17 not 18.

Rank example table

Employee	**Decoding Task score**	Rank
E16	100	1
E2	50	2
E1	50	2
E4	10	4
.	.	
.	.	
E3	4	16
E9	1	17
E13	1	17

The Different Contracts: You will have 18 contracts to choose from, numbered from 1-18 (Contract 1; Contract 2; . . . ; Contract 18). You can choose to select only one or all of the eighteen contracts. But, there is only one contract in each round that can give an Employee any earnings - **a contract only pays-off if its number is the same as the rank of that Employee!**

Distributing ECUs: In each round you will also choose how much ECUs to put on each of your chosen contracts. You have **19 ECUs to allocate** to your selected contact/s and there are **two requirements** that need to be fulfilled.

- **All the 19 ECUs must be used.**
- **One contract must get at least 1 ECU more compared to another.**

The "Most Preferred Contract": One Contract will be considered and called the **Most Preferred Contract**. It is the contract that an Employee has chosen to put the most ECUs on – the contract that fulfils the second requirement above.

ECU earnings of a Contract: The ECUs you allocate to a contract is called your **wage** (the pay per point). If a selected contract has the same number as your rank in that round, you will earn that **contract's wage times your score,** if that round is selected for payment.

Contract Selection Trial

Below you can try out different **Contract Selections** for a maximum of 4 minutes, or you can skip this by clicking "Next". The time starts counting down as soon as you make any choice.

The contract selection is done in two steps:

In **Step 1**, you select which of the 18 Contracts you would like and click "Select Contract(s)".
In **Step 2**, you will allocate the 19 ECUs (according to the two requirements) to your selected contracts of Step 1.
A number on the top right is showing how many ECUs you have left to decide about. If this figure becomes red and negative, you have allocated too many ECUs to the contract(s).

As soon as you would like to finalize your selection and allocation of ECUs, click "Submit Selection" and a summary of your choices will appear. If you would like to change your selected contracts of Step 1, click "Deselect Contract(s)".

In this trial, but not in the real rounds, you are able to select contracts as many times as you want during 4 trial minutes by clicking "New Trial Selection".

Select Contract(s) Trial:

Remaining trial time: 3:59

Contract 1 ○	Contract 2 ○	Contract 3 ○	Contract 4 ○	Contract 5 ○	Contract 6 ○
Contract 7 ○	Contract 8 ○	Contract 9 ○	Contract 10 ○	Contract 11 ○	Contract 12 ○
Contract 13 ○	Contract 14 ○	Contract 15 ○	Contract 16 ○	Contract 17 ○	Contract 18 ○

SELECT CONTRACT(S)

page 13 continuation,

Unfold once contracts are selected in Step 1.

Select Contract(s) Trial:

Remaining trial time: 3:23

- Insert a number between 1 and 19 for all of your Selected Contract(s).
- Remember to use all your 19 ECUs.
- Allocate (at least) one more ECU to one of the contract, compared to any other.

DESELECT CONTRACT(S)

ECUs left to allocate: **19 ECUs**

- [ ] **Contract 3:** You will be paid [X] ECUs/point **if your actual rank was 3**;
- [ ] **Contract 4:** You will be paid [X] ECUs/point **if your actual rank was 4**;
- [ ] **Contract 5:** You will be paid [X] ECUs/point **if your actual rank was 5**;
- [ ] **Contract 6:** You will be paid [X] ECUs/point **if your actual rank was 6**;

SUBMIT SELECTION

Summary: You selected [X] Contract(s) and your selection is valid. Your **Most Preferred Contract** was **Contract [Y]**, to which you allocated [YY] ECUs.

If you would like to try a new Contract Selection click “New Trial Selection".

NEW TRIAL SELECTION

Click "Next" to go to the last information about Selecting a Contract.

NEXT

Figure J.14: page 14

Summary information

When the Decoding Task and the Contract Selection of a round is finished the **Employee** will only get to know about their own **score** and knows their Selected Contract(s), while the **Principal** only get information about the **actual rank** of all its 3 Employees.

Below we show how this will look like for both the Employee and the Principal in a round 3 summary, which also includes information on the previous two rounds.

Example of an Employee's screen in Round 3

This is the summary of last round, including how you selected contracts.

Your score last round was 14.

ECU/Contract ("wage"[1])	Selected Contract(s)	
10 ECU	8	**Most Preferred Contract**
5 ECU	7	
3 ECU	5	
1 ECU	6	

In round 2: your **Most Preferred Contract** was 10.
In round 1: your **Most Preferred Contract** was 7.
*1) the **wage** only pays-off when the Rank and the Contract number equals.*

After the summary Employees are then asked to send some of the information to their principal. The **Principals** get a summary of the 3 Employees with the information below.

Example of a Principal's screen in round 3

Summary of round 3:
This is the summary of last round with the actual **ranks** of your 3 employees as well as a repetition of their round 2 and 1 rankings.

	Current Round 3	*Round 2*	*Round 1*
	Rank (actual)	Rank (actual)	Rank (actual)
Emp. 1	8	9	10
Emp. 2	3	2	3
Emp. 3	7	6	6

At the very end of the experiment, Employees and Principals will also learn their total earnings of the round that was randomly selected for payment of Part A. Next, we will ask you 5 questions about earnings and the contract selection, to make sure everybody understands. Click "Next" to go to the comprehension test.

Next

(a) Control

Summary information

When the Decoding Task and the Contract Selection of a round is finished the **Employee** will only get to know about their own **score** and knows their Selected Contract(s), while the **Principal** get information about the **actual rank, the**

Most Preferred Contract (the contract an Employee assigned the most ECUs), the **Difference** (between the actual rank and the Most Preferred Contract) and the **Direction** of the Difference (if the Difference was an under-, accurate or over-estimation) of all its 3 Employees.

Below we show how this will look like for both the Employee and the Principal in a round 3 summary, which also includes information on the previous two rounds.

Example of an Employee's screen in Round 3

This is the summary of last round, including how you selected contracts.

Your score last round was 14.

ECU/Contract ("wage"[1])	Selected Contract(s)	
10 ECU	8	**Most Preferred Contract**
5 ECU	7	
3 ECU	5	
1 ECU	6	

In round 2: your **Most Preferred Contract** was 10.
In round 1: your **Most Preferred Contract** was 7.
*1) the **wage** only pays-off when the Rank and the Contract number equals.*

After the summary Employees are then asked to send some of the information to their principal. The **Principals** get a summary of the 3 Employees with the information below.

Example of a Principal's screen in round 3

Summary of round 3:
This is the summary of last round with the actual **ranks** of your 3 employees as well as a repetition of their round 2 and 1 rankings.

In the table you can also see the information they have sent you, on the **Most Preferred Contract** (the contract given the most ECUs); as well as the **Difference** (between actual rank and the Most Preferred Contract) and the **Direction** of that difference with $+$, $-$ and ± 0 indications for an over-, under- and an accurate- estimation, respectively).

	Current Round 3			
	Rank (actual)	Contract (preferred)	Diff. (Rank - MPC)	Direction (of Diff.)
Emp. 1	8	5	$+3$	Overestimation
Emp. 2	3	2	$+1$	Overestimation
Emp. 3	7	9	-2	Underestimation

	Round 2				Round 1			
	Rank	MPC	Diff.	Dir.	Rank	MPC	Diff.	Dir.
Emp. 1	9	14	-5	Underest.	10	7	$+3$	Overest.
Emp. 2	2	2	± 0	Accurate est.	3	3	± 0	Accurate est.
Emp. 3	6	4	$+2$	Overest.	6	8	-2	Underest.

At the very end of the experiment, Employees and Principals will also learn their total earnings of the round that was randomly selected for payment of Part A. Next, we will ask you 5 questions about earnings and the contract selection, to make sure everybody understands. Click "Next" to go to the comprehension test.

Next

(b) Public Treatment

Figure J.15: page 15

Comprehension test

Here is an example of a Decoding Task and a Contract Selection outcome. In the example, there are only 4 instead of 18 Employees, called E1, E2, E3 and E18. You need to answer all the 5 questions correctly to proceed.

Example outcome:

	Score (total points)	**Rank** (actual)	**Selected Contract(s)** (allocated ECUs)	**Wage** (earning/point)	**Earnings** (total pay)
E1	7	3	Contract 6 (10 ECUs); Contract 5 (5 ECUs); Contract 4 (4 ECUs)	0 ECUs	0 ECUs
E2	6	4	Contract 5 (8 ECUs); Contract 4 (6 ECUs); Contract 3 (4 ECUs); Contract 6 (1 ECUs)	? ECUs	? ECUs
E3	9	1	Contract 2 (15 ECUs); Contract 1 (4 ECUs)	4 ECUs	? ECUs
E18	8	2	Contract 3 (9 ECUs); Contract 1 (5 ECUs); Contract 2 (5 ECUs)	5 ECUs	? ECUs

1. **Which is the "Most Preferred Contract" of E1?**
   Contract 3 ◯ Contract 4 ◯ Contract 5 ◯ Contract 6 ◯
2. **Which is the "Most Preferred Contract" of E2?**
   Contract 3 ◯ Contract 4 ◯ Contract 5 ◯ Contract 6 ◯
3. **What is the wage of E2?**
   4 ECUs ◯ 6 ECUs ◯ 8 ECUs ◯ 1 ECU ◯
4. **How much will E3 earn in total (for this example round)?**
   9 ECUs ◯ 135 ECUs ◯ 0 ECUs ◯ 36 ECU◯
5. **Who will earn the most of E3 and E18?**
   E3 will earn 4 ECUs more than E18 ◯
   E18 will earn 4 ECUs more than E3 ◯
   They will earn the same ◯
   None of them will earn any ECUs ◯

Answer all the 5 questions and click "Submit". If an answer is incorrect, it is marked and needs to be changed.

SUBMIT

`Indicate where wrong. Cannot pass without all correct.`

Figure J.16: page 16

Your role

In the beginning of the experiment you were randomly assigned to the role: [**Employee/Principal!**]
Note: You will keep the **same role** and **group** (of 1 Principal and 3 Employees) **for Part A**, but not later in the experiment.

Click "Next" to select Contract(s)

NEXT

Figure J.17: page 17

Contract Selection

In round [1] **you scored [X]** in the Decoding Task. You will now select a contract to be paying you for this work!

Your earnings depend on your choices! Likewise, will your Principal earn a third of what you earn plus a third of the other two Employees' payment that your Principal has been matched with.

Remember that the Contract only pays-off **if its number is the same as your actual rank!**

You will first have 18 contracts to select from and then you will have 19 ECUs to distribute over these selected Contract(s).

Information: After you have selected your contract(s) we will ask you to submit information about your actual **Rank** to your Principal.

You will get a summary reminder of your **selected Contract(s)** and your **score** as described before.

Select your Contract(s) below in two steps:

[SELECTS CONTRACT(S) AS OUTLINED IN Figure J.13]

`No time-restriction`

SUBMIT SELECTION

NEXT

(a) Control

Contract Selection

In round [1] **you scored [X]** in the Decoding Task. You will now select a contract to be paying you for this work!

Your earnings depend on your choices! Likewise, will your Principal earn a third of what you earn plus a third of the other two Employees' payment that your Principal has been matched with.

Remember that the Contract only pays-off **if its number is the same as your actual rank!**

You will first have 18 contracts to select from and then you will have 19 ECUs to distribute over these selected Contract(s).

Information: After you have selected your contract(s) we will ask you to submit information about your actual **Rank** and your selected **Most Preferred Contract** as well as the **Difference** (Rank-MPC) between those and the **Direction** of the difference to your Principal.

You will get a summary reminder of your **selected Contract(s)** and your **score** as described before.

Note that you will not know your actual **Rank** and neither the **Difference** nor the **Direction** of the difference, when you submit it to your Principal.

Select your Contract(s) below in two steps:

[SELECTS CONTRACT(S) AS OUTLINED IN Figure J.13]

`No time-restriction`

SUBMIT SELECTION

NEXT

(b) Public Treatment

Figure J.18: page 18

[SUMMARY OF ROUND 1/2/3 as outlined in upper part of Figure J.14a and J.14b]

Figure J.19: page 19

Submitting information

Information: Below there is **information** that we ask you to **send to your Principal**, together with the information that you have finished round 2.

The principal will get information about your **actual rank**. To finish click "Submit information to my Principal".

Submission to my Principal

I have now finished round 2 and I am submitting information about my **rank**.

Submit information to my Principal

(a) Control

Submitting information

Information: Below there is **information** that we ask you to **send to your Principal**, together with the information that you have finished round 2.

The principal will get information about your **actual rank** and we ask you to fill in your "Most Preferred Contract" of this round in the box. To finish click "Submit information to my Principal".

Submission to my Principal

I have now finished round 2 and I am submitting information about my **rank**.

: ______ as my **Most Preferred Contract**.

Submit information to my Principal

(b) Public Treatment

Figure J.20: page 20

End of Round

You have now finished round [1/2/3] and you will now be waiting for the others to finish this round as well.

Click "Next" to proceed.

Next

Other Experimental Parts

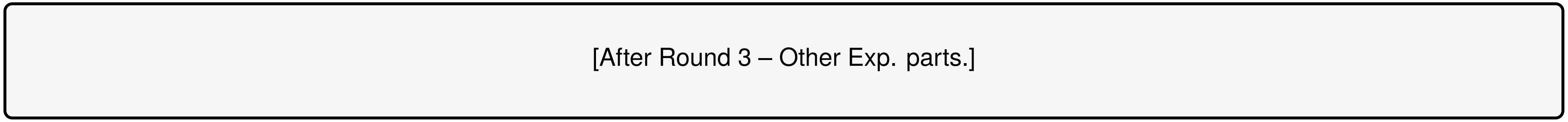


Figure J.21: page 27

Figure J.22: page 28

Payment Information

In the table below you can find your final pay for all parts of the experiment that could give earnings.

Earnings in Part A

In Part A a random draw decided that **round [1/2/3]** will be used for payment.

Your **Contract(s) Selection**, **rank** and **score** in Round [1/2/3] plus the **random pay** of 0, 1 or 2 times your score will decide your earnings.

Earnings in Part B

In the lottery, number [X] was randomly drawn for payment. You selected [XX].

Summary table of your total payment

Show-up fee	**5 €**
Experiment Part:	**Earnings**
Part A (main part)	[XA] ECUs
Part A (score guess)	[XAE] ECUs
Part B (Lottery)	[XB] ECUs
Total: [T]	**5 € + X[A+ AE or (AP + AP2) + C] ECUs**

Your total payment is [T], which includes the show-up fee of 5 € and your earning of the experiment of [T-5] € €

The experiment is now finished and we want to thank you for your participation!

Important payment information:

[. . .]

(a) Control

Payment Information

In the table below you can find your final pay for all parts of the experiment that could give earnings.

Earnings in Part A

In Part A a random draw decided that **round [1/2/3]** will be used for payment.

Your **Contract(s) Selection**, **rank** and **score** in Round [1/2/3] will decide your earnings.

Earnings in Part B

In the lottery, number [X] was randomly drawn for payment. You selected [XX].

Summary table of your total payment

Show-up fee	**5 €**
Experiment Part:	**Earnings**
Part A (main part)	[XA] ECUs
Part A (score guess)	[XAE] ECUs
Part B (Lottery)	[XB] ECUs
Total: [T]	**5 € + X[A+ AE or (AP + AP2) + C] ECUs**

Your total payment is [T], which includes the show-up fee of 5 € and your earning of the experiment of [T-5] € €

The experiment is now finished and we want to thank you for your participation!

Important payment information:

[. . .]

(b) Public Treatment

End of Appendices.