

(Over)claiming Credit in Collaborations *

Marissa Lepper Jonas Mueller-Gastell Stephanie W. Wang

Preliminary Draft

August, 2023

Abstract

Claiming credit for contributions in collaborative work is an essential part of many workplace decisions from hiring to promotions. We experimentally test whether people accurately assess their contribution in real-effort group tasks and find evidence of systematic overclaiming. We explore factors that could exacerbate or reduce overclaiming such as imperfect memory, the degree of ambiguity about contributions, and social pressure. Allowing teammates to discuss respective contributions prior to making claims reduces overclaiming.

1 Introduction

Many organizations, including companies and universities, rely heavily on work completed by teams. However, it can be difficult to monitor each individual's contribution to the group's shared goal. It is therefore common for hiring and promotion processes to rely on self-reported contributions as a proxy. Examples of this include job candidates frequently being asked to estimate their contributions to recent projects and promotion decisions hinging on employees' claims about contributions to past team accomplishments, such as reports about client acquisition and case involvement at law firms.

We explore the accuracy of self assessments about contributions to team work. Moreover, we study factors that may influence the observed gap between credit claimed and credit

*Lepper: University of Pittsburgh, Department of Economics, Wesley W. Posvar Hall, Pittsburgh, PA 15260; email MAL303@pitt.edu. Mueller-Gastell: Stanford University, jonasmg@stanford.edu Wang: University of Pittsburgh, Department of Economics, Wesley W Posvar Hall, Pittsburgh, PA 15260, swwang@pitt.edu

ascribed by neutral third-party observers. Specifically, participants in our experiment collaborate on real effort tasks such as solving a riddle in teams of three. Communication within groups can improve task performance and individual compensation depends on one randomly selected answer per group, thus encouraging within-group collaboration. The solution to the task is unique and easily identifiable, allowing groups to work towards a common known goal. Participants then make incentivized claims about their own and their teammates' contributions to the task. We find robust evidence of overclaiming credit: Participants ascribe more credit to themselves compared to average objective rating they received 70 percent of the time. Fifty-three percent of participants assign over half of the credit to themselves, while only 25 percent pass this threshold in objective ratings. Finally, we find that over a third of participants overestimate their relative contribution to the group.

We explore potential mechanisms underlying this overclaiming phenomenon with several treatments. In the *Public* treatment, participants now see their teammates' credit claims and are made aware this prior to making claims. Additionally, team members discuss claims in a chat before making them in some of the *Public* claims are made following a small chat in most sessions. This allows us to better understand the social norms surrounding credit claiming and how social pressure may affect the extent of overclaiming. We find that overclaiming is reduced in when contribution claims are made public, especially when those claims follow a chat with team members about contributions from each member.

Self-serving bias and selective recall may exacerbate overclaiming. We study this in the *Delay* treatment by eliciting beliefs three days after the task. Similar to the *Baseline* treatment, we find persistent overclaiming. This is of particular concern due to hiring and promotion decisions often relying on delayed self-report contribution claims over past events.

2 Riddles Task

Groups of three collaborate via chat to solve a riddle.¹ Each riddle has an unique answer which, while not immediately obvious, is self-evident once found. Although each participant submits an answer, compensation for the entire group depends on one randomly selected answer. This setup mirrors the common workplace situation in which teammates work toward one common goal, but each member's level of contribution is subjective.

participants are incentivized to accurately assign their contribution as well as the con-

¹We use three riddles, each adapted from the "stumpers" use by Bar-Hillel, Noah and Shane (2018). See the for additional details.

tributions of the other two group members (the percentages must add up to 100). mTurk participants in separate sessions then act as third-party raters and do the same after reading the group chats and seeing the submitted answers.² We use the average of the three to five ratings gathered for each participant as our *objective* metric of contribution (all results are robust to using the median rating instead).

2.1 Session Overview

The *Baseline* study was run in November 2018, March 2019, and April 2019 with 291 participants from Amazon Mechanical Turk (35.42 years old on average and 46.7 percent female).³ Sessions lasted around 30 minutes with an average payment of \$7.05, comprised of a \$0.50 HIT fee, a completion payment of \$3.50, and an average bonus of \$3.05.

Participants first fill out a short demographic form, which is used to ensure a gender-balanced sample by screening out 25 percent of participants who identify as male.⁴ Those who continue forward next answer a science quiz comprising of 20 multiple-choice science questions worth 5 cents each. They then wait until a group of three can be formed.⁵ Once grouped, participants see their teammates' demographic information, including gendered nicknames used to identify each individual within the chat. Groups spend at least three but no more than eight minutes discussing the riddle and at least one minute on each rating page, ensuring thoughtful estimations of contribution. Each group member receives a \$2 bonus if the selected answer is correct, regardless of their own answer. Additionally, we pay participants up to \$1 each for their own credit claim and that of one randomly-selected teammate.⁶

Finally, participants answer a series of additional incentivized questions. These include measures of overconfidence, such as guessing their science quiz performance, a risk attitude elicitation, and an additional demographic questionnaire. Bonus payments are distributed three days after each session to allow the third-party ratings to be collected and assigned.

²Additional information about rating sessions is available in the Appendix.

³Workers outside of the US, with fewer than 1,000 HITs, or with an approval rating under 97 percent were not eligible to participate.

⁴These participants, who are not included in summary statistics, were paid the HIT payment and an additional \$0.50 bonus.

⁵Anyone not paired within 15 minutes can opt out, receiving the HIT payment and a \$1.50 waiting payment.

⁶We use a quadratic scoring rule with accuracy determined by the average of their third-party ratings of contribution.

2.2 Results

We find systematic overclaiming compared to objective third-party ratings. Participants overclaim both the amount of credit they should receive (shown in Section 2.2.1) as well as the relative rank of their contribution compared to their teammates' (shown in Section 2.2.2). In 2.3, we establish that overclaiming is a behavioral phenomenon distinct from overconfidence. Finally, in Section 2.4, we explore mechanisms that could exacerbate or reduce overclaiming: delayed recall and public discussions about credit with other team members.

2.2.1 Overclaiming of Absolute Contribution

Table 1 compares self-assigned contribution claims to those assigned by third-party raters. Overclaiming happens at both the extensive margin, with self-reported contributions exceeding those of the third-party raters 70 percent of the time compared to the average rating and 65 percent compared to the median, and the intensive margin. On average, participants claim 50.97 percent of the credit compared to the approximately 33.3 percent we would expect without overclaiming mechanically by virtue of three people dividing 100 percent credit.⁷ We find systematic overclaiming, though marginally mitigated, even when we only look at claims made by the 75 percent of participants who were successful at solving the riddle, thus removing any difficulty that may come from assigning contribution to an unsuccessful task. Successful participants, on average, claim 50.52 percent of the credit but are only given an average rating of 35.28 percent ($t = 7.895$, $p < 0.01$).

Figure 1 shows a scatterplot with self-reported credit on the y-axis and assigned credit on the x-axis with histograms of each distribution. This allows us to explore how distributional differences between self- and assigned-credit claims. This graph illustrates a handful of important results. First, overclaiming is more common than under-claiming at all levels of assigned credit. Second, there are large clusters for self-assigned ratings at both full and no credit, and smaller clusters around half- and one-third-credit. Third-party ratings, on the other hand, are lower across the board and display less round-number attraction.

The differences in the distributions of self- and third-party-ratings is best seen by com-

⁷Although each participant must split 100 percent of credit between themselves and their group members, the outcome variable of *own* contribution is not mutually exclusive within a group. Accurate assessments by all group members, however, would still sum to 100. An average contribution higher than one-third therefore shows population-level overclaiming. Mechanically, the average third-party rating does sum to approximately one-third as each triplet of participants is collected and forced to be mutually exclusive. The small deviation is caused by the handful of participants dropped who did not complete the study.

Table 1: Comparison between Self-Assigned and Third-Party Assigned Contribution

	Self	Third-Party	Difference
Average Amount Assigned	50.97 (2.04)	32.78 (1.58)	18.18*** (2.58)
(Self > Third Party)	70.44%*** (2.67)	-	-
Assigned Above 50	53.61% (2.93)	25.77% (2.57)	27.83*** (3.89)
Assigned Above 33	65.63% (2.79)	37.11% (2.83)	28.52*** (3.98)
<i>N</i>	291	291	291

Note: Results of t-tests comparing self-assigned and third-party assigned contribution. The last three variables represent percentages of the sample who claim a higher contribution than they are assigned (*Self > Third Party*), who claim/are assigned a contribution amount over half (*Assigned Above 50*), and who claim/are assigned a contribution amount over one-third (*Assigned Above 33*). Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

paring how many participants reach a certain threshold of credit, such as half or one-third of credit. Looking at self-ratings, over half of participants claim they deserve at least half of the credit. However, only around a quarter actually receive an assignment over 50 percent. Similarly, there is a gap of 28 percentage points between the proportion of participants assigning themselves over a third of the credit compared to the proportion who reach this threshold using third-party ratings.

2.2.2 Overclaiming of Relative Contribution

We now explore overclaiming in terms of relative contribution, thus removing differences which move the absolute amount contributed but preserve rank between group members. For example, the subjective nature of the ranking may allow for motivated beliefs, pushing self-assessment above the true amount. However, these may be constrained by a correct assessment of relative contribution, where participants cannot move their belief of their ranking above those they assess as contributing more. Alternatively, anchoring points may differ between participant and raters, with participants potentially anchoring at 100 percent and raters at one-third. Relative contribution therefore serves as a stricter test of overclaiming.

We calculate relative rank by counting who in the team is assigned a *strictly* higher absolute contribution, creating an upper-bound of rank. Mechanically, around one-third of participants are assigned each rank. As illustrated in Figure 2, overclaiming in terms of rank

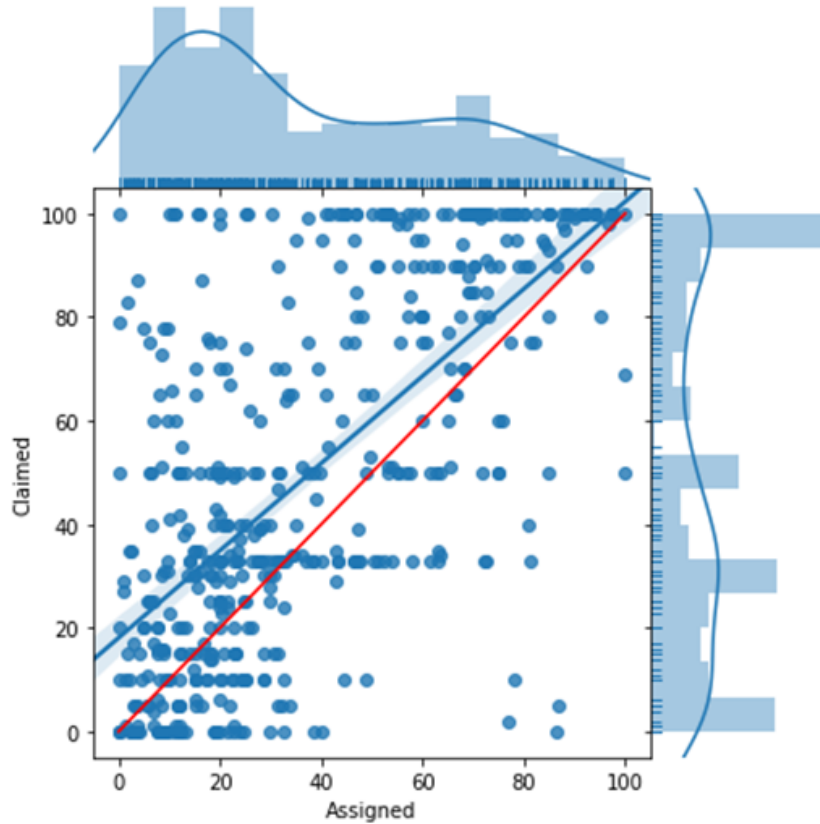


Figure 1: Scatterplot of Self-Assigned and Third-Party Contribution Rating

Note: Scatterplot showing self-reported credit on the y-axis against assigned credit on the x-axis and the results of a linear regression of claim on assigned credit with 95 percent confidence bands. The 45-degree line (in red) corresponds to claiming the true contribution, with any points above or below representing over- or under-claiming, respectively. The histograms on the axes show the overall distribution of each the two ratings of contribution: own ratings on the vertical histogram and third-party ratings on the horizontal histogram. Sample restricted to those in Baseline Riddle Experiment who completed the entire survey.

is present, although attenuated, happening with 36 percent of participants.⁸ Overclaiming is mostly used to claim weakly the most contribution, which happens in over two-thirds of the cases and leads to 58.76 percent of participants claiming this rank. However, overclaiming is also used to avoid the weakly worst ranking by reporting the second-highest contribution, with almost a third of those claiming this rank not truly deserving it.

Participants who assess their relative contribution accurately still overclaim in absolute amounts, average 21.4 percentage points above the their third-party rating. However, the almost half of participants (47 percent) who overclaim in both dimensions do so by a far

⁸If we instead use weakly higher contributions to determine ranks, thus measuring a lower-bound of each participant's ranking, overclaiming drops to 30 percent.

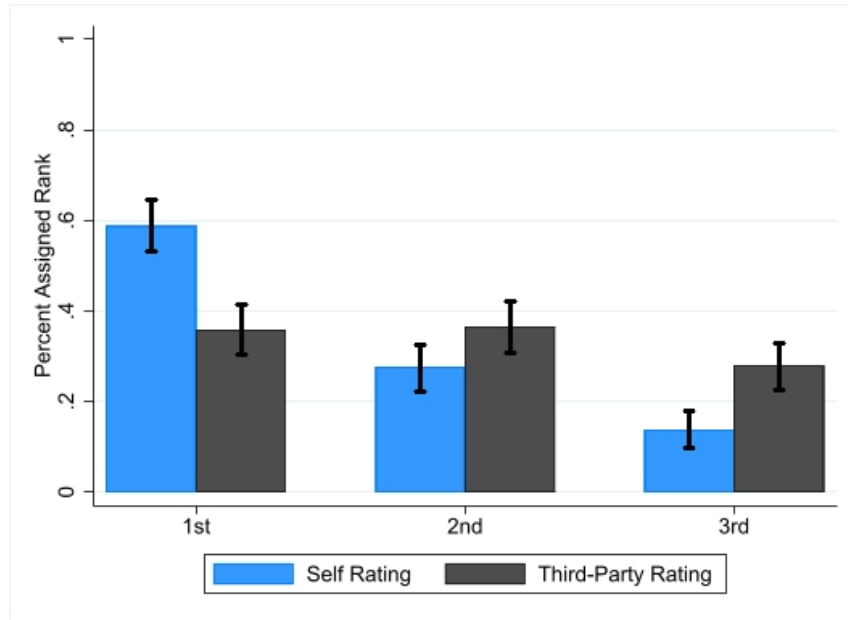


Figure 2: Comparison of Rank between Self- and Third-Party-Rating

Note: Assigned rank is the implied rank determined by counting the number of team members given strictly higher contribution amounts. Includes 95 percent confidence intervals.

higher amount, averaging an absolute overclaiming amount of 42.6 percentage points ($t = 7.843$, $p < 0.01$). This suggests that while overclaiming is still present, relative ranking can act as a bound on how large of a deviation from true contribution participants claim. Moreover, eliciting ranks of contribution rather than absolute numbers will create more accurate assessments.

2.3 Overconfidence

We next establish that overclaiming is a behavioral phenomenon distinct from overconfidence using the science quiz. Table 1 shows summary statistics split by gender. Women perform marginally better than men on the quiz, but men *believe* they perform significantly better. Moreover, significantly more men guess their quiz score is higher than it actually is. Table 3 shows the results of a linear regression of actual quiz score and being a woman on guessed quiz score. Women, on average, are 5 percentage points closer to their true score than men. These results replicate the well-established fact that overconfidence is prevalent in general, and more so for men (Barber and Odean, 2001; Niederle and Vesterlund, 2007; Hügelschäfer and Achtziger, 2014).

If overclaiming is a result of overconfidence we expect two results will hold: (I) people who

Table 2: Quiz Performance and Overconfidence Measures by Gender

	Male	Female	Difference	se
Quiz Score	25.81%	28.57%	-2.76	2.02
Score Guess	43.87%	39.88%	3.98*	2.15
% Overconfident	79.35	68.38	10.97**	5.12
<i>N</i>	155	136	291	

Note: Summary statistics and results of t-tests using the 20-question science quiz. *Quiz Score* is the true percent of questions answered correctly, while *Score Guess* is their guess. *% Overconfident* is the percent guessing score above their actual one. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: OLS Estimates of Quiz Performance Belief

	Performance Belief
Score	0.403*** (0.0569)
Female	-5.189*** (1.982)
Constant	33.56*** (1.929)
<i>N</i>	278

Note: Results of a linear regression of actual quiz performance (*Score*, measured as a percent of quiz questions) and gender (*Female*, an indicator variable for being female) on beliefs about the percent of questions answered correctly. Standard errors in parentheses clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

are overconfident will overclaim more; and (II) men will overclaim more than women, as they are more overconfident. Table 4 explores the impact of gender and overconfidence on various measures of overclaiming within the *Baseline* treatment. Rows (1) – (3) use the dependent variable of the difference between self-report and third-party assessment. Columns (4) – (6) look at the binary variable of claiming a higher absolute amount of credit. Columns (7) – (9) look at the binary variable indicating if a participant overclaims their relative rank. *Overconfident* is an indicator variable for guessing a higher score than was actually earned on the science quiz and *Female* is an indicator variable for being female.

Interestingly, controlling for third-party ratings does not significantly impact whether someone overclaims or not in absolute terms, but rather how *much* they overclaim. However, no specification shows any association between either any measure of overconfidence or gender and overclaiming. Thus, overclaiming behavior is not a direct result of being overconfident, but a separate phenomena that affects both genders. De-biasing attempts must therefore

treat it as a distinct behavioral bias.

2.4 Mechanisms

We run two additional treatments to identify features at the elicitation stage which affect overclaiming. First, we explore the effect of delaying the assessment in the *Delay* treatment. Second, in the *Public* treatment, groups are given the opportunity to discuss contributions with each other, which are thereafter shared with other group members. These serve to test the robustness of overclaiming in settings similar to those used in the workplace which can also be used to adjust overclaiming amounts, suggest mechanisms which lead to overclaiming, and create direct and actionable policy recommendations for accurate elicitation of self-contribution in the workplace.

Hiring and promotion decisions often rely on self-reports over past events, which may attenuate or exaggerate credit claims by introducing a memory component in *Delay*. Time delays have been shown to exacerbate motivated reasoning in other domains, such as assessing one’s own mental abilities (Zimmermann, 2020). Overclaiming increasing after a time delay suggests that similar motivated reasoning forces are at play. The *Public* treatment introduces a social component by adding discussion and transparency. This further interaction may moderate overclaiming through challenges to high claims by group members and an increase in the salience of others’ contributions. Further, participants may want to avoid appearing greedy or confrontational, or be identified as outright lying to their team members (consider, e.g. Abeler, Nosenzo and Raymond, 2019). On the other hand, status concerns, boasting, and differences in negotiating ability may all exacerbate overclaiming.

2.4.1 Method

Both de-biasing treatments, the sample sizes of which are reported in Table 5, used the Riddle task described in Section ?? and were run on mTurk in April 2019 (*Public*) and April 2020 (*Delay*).

Compared to *Baseline*, *Delay* differs only in the timing of the belief elicitation, which occurs three days after the task.⁹ *Public* differs by giving participants the opportunity to chat with their teammates prior to assigning contribution and then displaying elicited contributions to the other group members. Outside of an additional completion payment

⁹In *Baseline*, participants are shown their group’s chat transcript while assigning contributions. In *Delay*, participants are instead shown it at the end of the first session and informed that the follow-up will ask about how their team worked.

Table 4: OLS Estimates of the Impact of Gender and Overconfidence on Overclaiming in Baseline

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Δ	Δ	Δ	Binary Overclaim	Binary Overclaim	Binary Overclaim	Rank Overclaim	Rank Overclaim	Rank Overclaim
Overconfident	2.059 (3.835)		0.628 (4.173)	-0.00296 (0.0613)		-0.0163 (0.0711)	-0.0169 (0.0649)		-0.0381 (0.0651)
Gender		0.0630 (3.305)	1.541 (3.499)		0.0165 (0.0537)	0.0150 (0.0594)		-0.0838 (0.0562)	-0.0209 (0.0545)
Third-Party Rating			-0.220** (0.0580)			-0.000363 (0.00109)			-0.00756*** (0.000765)
Answer Correct			-9.819** (4.340)			-0.127* (0.0693)			-0.134* (0.0790)
Quiz Score			-14.13 (11.58)			-0.219 (0.202)			-0.134 (0.180)
Constant	16.66*** (3.320)	18.16*** (2.409)	30.67** (11.86)	0.707*** (0.0528)	0.697*** (0.0370)	0.705*** (0.209)	0.373*** (0.0560)	0.400*** (0.0395)	0.830*** (0.202)
<i>N</i>	291	291	278	291	291	278	291	291	278
Controls	No	No	Yes	No	No	Yes	No	No	Yes

Note: Standard errors in parentheses clustered at the individual level. Controls include riddle fixed effects and demographic information such as age, educational attainment, marital status, income, ethnicity, and occupation. The outcome variables are: Δ *Overclaim*, the difference between self-rated and third-party rated contribution; *Binary Overclaim*, a binary variable which is one for those with self-rated contributions above their third-party ratings; *Rank Overclaim*, a binary variable which is one for those with a higher relative self-rated rank than that assigned by third-party raters. *Overconfident* is an indicator variable for having high beliefs than true score in the quiz. *Gender* is an indicator variable for being female, while *Answer Correct* is an indicator variable for being successful in the riddle. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Sample Sizes for the Riddle Task Treatments

	No Delay	Delay
Private	291	195
Public	290	-

Note: Private-No Delay is the *Baseline* treatment, Private - Delay is *Delay* and Public-No Delay is *Public*. Sample size only reports participants who were not screened out in the demographic questionnaire.

added in *Delay*, pecuniary benefits remain constant for both treatments.¹⁰ We can therefore attribute any differences causally to the change in elicitation methods, creating costless policies that impact veracity of claims.

2.4.2 Results

We find overclaiming across all treatments. Although participants complete the same task, performance varies between treatments, with success in solving the riddle being more common in *Delay* and less common in *Public*. Average third-party ratings reflect this: participants in *Delay* receive more credit and those in *Public* receive less. Table 6 summarizes contribution claims and over-claims for the three treatments. *Delay* does not influence overclaiming: similar to *Baseline*, around 70 percent of participants claim a higher contribution than they were assigned, with an average amount of overclaiming around 18 percentage points. Overclaiming relative contribution is marginally less prevalent, with a little over a third of the sample claiming a higher rank. *Public*, however, does mitigate overclaiming. Only 62 percent of participants assign themselves a higher contribution than they are rated by an average of only 12.5 percentage points. Although discussing and sharing claims lowers *absolute* overclaiming, relative overclaiming does not appear to be impacted.

The regression analysis in Table 7 further explores the impact of treatments on various measures of overclaiming, including the absolute amount overclaimed on both the intensive and extensive margin and relative overclaiming. We confirm the attenuation of absolute overclaiming seen in *Public*, however, once we control for factors such as demographics, correctly solving the riddle, and overconfidence about performance on the science quiz, the reduction in the percent of participants overclaiming is no longer significant. These results seem to indicate that the treatment can reduce overclaiming on the intensive margin, which

¹⁰To reduce attrition, participants receive an additional \$1.00 completion bonus following the first session; all other bonuses and a \$3.00 completion bonus are withheld until completion of the follow-up survey. Rating happens between the sessions, allowing participants to receive their bonus the same day as the follow-up survey. *Baseline* also involves a delay in payment, minimizes potential selection effects.

Table 6: Overclaiming Statistics for all Treatments in the Riddle Tasks

	Baseline	Delay	Public
Claimed	50.97	53.95	43.34
Third-Party Assigned	32.78	35.16	30.62
Correct Answer	70.79%	78.87%	57.58%
Overclaim Absolute	70.44%	70.17%	62.07%
Overclaim Rank	36.08%	34.50%	36.21%

Note: *Claimed* is self-assigned contribution and *Third-Party Assigned* is the average of the third-party ratings. The last three variables represent percentages of the sample who answer the riddle correctly individually (*Correct Answer*), who claim a higher contribution than they are assigned (*Overclaim Absolute*), and who claim a higher rank than they are assigned (*Overclaim Rank*). Does not include participants who do not submit contribution claims.

is where it is most prevalent, but does not move other starker measures.

Finally, Table 8 further explores the role of overconfidence in overclaiming. First, there is no impact of gender, as we would expect if overconfidence was driving our results.¹¹ There is also no overall role of overconfidence, which further confirms that this is a distinct behavioral bias. Interestingly, there is a significant interaction between overconfidence and relative overclaiming in *Public*: those who continue to claim too high of a rank, even after discussing with their teammates and with the understanding their claims will be shown, tend to be those who are overconfident in the science quiz. This shows that there may be residual overclaiming caused by overconfidence even after a debiasing treatment.

3 Math Task

3.1 Method

The Math Task was run in July and September 2020 and February through May 2021, amassing a total of 722 participants. We ran three treatments (*Baseline*, *Delay*, and *Public*) and list the per-treatment sample sizes in Table 9. Sessions, which varied in size between 12 and 24 participants, were run virtually in the Pittsburgh Experimental Economics Laboratory (PEEL) over Zoom following the virtual procedures laid out in Danz, Gupta, Lepper, Vesterlund and Winichakul (2021). Although the average duration of the Math and Riddle task were identical, we modified average payment amounts to be in line with the standards in

¹¹In the Appendix, we show that there is no significant interaction of either treatment with gender either.

Table 7: OLS Estimate of Treatment Effects in the Riddle Task

	(1)	(2)	(3)	(4)	(5)	(6)
	Δ	Δ	Absolute	Absolute	Relative	Relative
	Claim	Claim	Overclaim	Overclaim	Overclaim	Overclaim
Public	-7.189*** (2.338)	-4.693** (2.325)	-0.086** (0.041)	-0.060 (0.040)	-0.018 (0.038)	-0.027 (0.037)
Delay	0.308 (2.702)	1.716 (2.732)	-0.007 (0.046)	-0.009 (0.045)	-0.000 (0.043)	0.006 (0.042)
Third-Party	-0.297*** (0.039)	-0.320*** (0.039)	-0.001 (0.001)	-0.001* (0.001)	-0.008*** (0.000)	-0.008*** (0.000)
Constant	32.515*** (3.542)	25.096*** (5.487)	0.749*** (0.062)	0.611*** (0.099)	0.646*** (0.059)	0.723*** (0.092)
<i>N</i>	750	722	750	722	750	722
Controls	No	Yes	No	Yes	No	Yes

Note: Standard errors in parentheses clustered at individual level. Controls include demographic information such as riddle fixed effects, gender, marital status, income, ethnicity, and occupation. Outcome variables are: Δ *Overclaim*, the difference between self-rated and third-party rated contribution; *Absolute Overclaim*, a binary variable which is one for those with self-rated contributions above their third-party ratings; *Relative Overclaim*, a binary variable which is one for those with a higher relative self-rated rank than that assigned by third-party raters. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

the new experimental setting.¹² The percentage of male-identifying subjects is 47 (*Delay*), 50.5 percent (*Public - Chat*), and 37.2 percent (*Public - No Chat*), and the average age is 20 years old in all treatments.

3.1.1 Math Task Overview

Similar to the Riddle treatments, participants work on a collaborative real-effort task in groups of three. However, the task now involves adding nine integers ranging from -3 to 3 (excluding 0), presented in a square. One player is the active player and can submit an answer for the square, which is visible on all group members' screens. Each correct answer increases the payment of all members in the group by \$0.25. After an active player submits an answer, regardless of its accuracy, one of the remaining two reserve players becomes the new active player in a set order and can submit an answer for a new square. Reserve players

¹²We increased the base pay to \$6 but kept incentivized payments similar. Average payment for *Baseline* and *Public* were around \$16. *Delay* had an additional completion bonus to reduce attrition, which increased the average payment to \$20. Moreover, all incentivized bonuses were only paid after the completion of the entire study. We exclude anyone who did not return in the *Delay* treatment, but results are robust to including them, as shown in the Appendix.

Table 8: OLS Estimate of Treatment Effects in the Riddle Task

	(1)	(2)	(3)	(4)	(5)	(6)
	Δ	Δ	Absolute	Absolute	Relative	Relative
	Claim	Claim	Overclaim	Overclaim	Overclaim	Overclaim
Public	-5.596** (2.363)	-7.367* (4.055)	-0.065 (0.041)	-0.105 (0.069)	-0.0345 (0.039)	-0.115* (0.062)
Delay	0.746 (2.744)	1.287 (4.402)	-0.014 (0.046)	-0.001 (0.075)	-0.002 (0.044)	-0.082 (0.069)
Female	0.390 (2.050)	0.607 (2.086)	0.036 (0.035)	0.040 (0.035)	0.004 (0.032)	0.002 (0.032)
Third-Party Rating	-0.317*** (0.040)	-0.311*** (0.039)	-0.001* (0.001)	-0.001 (0.001)	-0.008*** (0.000)	-0.007*** (0.000)
Overconfident		1.304 (3.840)		0.007 (0.063)		-0.034 (0.060)
PublicXOverconfident		3.714 (5.051)		0.077 (0.086)		0.138* (0.079)
DelayXOverconfident		-0.523 (5.778)		-0.020 (0.095)		0.142 (0.088)
Constant	28.87*** (6.088)	27.58*** (6.779)	0.630*** (0.112)	0.619*** (0.121)	0.757*** (0.104)	0.779*** (0.113)
<i>N</i>	722	722	722	722	722	722
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Note: Standard errors in parentheses clustered at individual level. Controls include demographic information such as riddle fixed effects, gender, marital status, income, ethnicity, and occupation. Outcome variables are: Δ *Overclaim*, the difference between self-rated and third-party rated contribution; *Binary Overclaim*, a binary variable which is one for those with self-rated contributions above their third-party ratings; *Rank Overclaim*, a binary variable which is one for those with a higher relative self-rated rank than that assigned by third-party raters. *Female* indicates identifying as female and *Overconfident* is having inaccurately high beliefs about science quiz performance. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

can vote to skip the current active player. If both reserve players do so, the next reserve player becomes the active player and can submit an answer for that square. Groups share a time limit and work towards the shared goal of increasing the total earnings of the group, incentivizing participants to skip over slower or less accurate participants in order to submit the answer themselves.

Participants next make claims about contribution for themselves and their teammates. We use the percentage contributed to the group score as a proxy for actual contribution.¹³ We find that percentage of squares solved explains 78 percent of the variance in the third-party raters' contribution assignments, and the R2 is only increased to 0.81 by including all

¹³Third-party raters were only used in the *Delay* treatment.

Table 9: Sample Size for Math Task Treatments

	No Delay		Delay
	Chat	No Chat	No Chat
Private	-	-	277
Public	168	51	-

Note: Table shows effective sample size, only including participants who complete all experimental measures.

other performance variables (such as number of squares solved, total time taken, ...). Thus, the percentage of squares solved is use as an objective proxy for contribution.

3.1.2 Session Overview

Participants first complete the same demographic survey and science quiz as in the Riddle task. Participants then solve squares individually, receiving \$0.10 per correctly solved square, in a two minute practice round. They next see demographic information about their teammates, who are identified to each other using assigned same-gender nicknames. After the group task participants give their contribution claims and answer several other incentivized questions, including guessing their score in the science quiz guessing how many squares their group submitted in total, and what percentage they contributed to that, and a risk preference elicitation.

The *Delay* and *Public* treatments are similar to their counterparts in the Riddle task. Participants in *Delay* complete the belief elicitation three days after the task. Two types of *Public* are run: *Chat* vs *No-Chat*. We share each participants' contribution claim with their group in both; however, in the *No-Chat* treatment, participants are not given the chance to discuss contributions with their teammates. We can then disentangle the effects of the discussion from the effects of publicly revealing claims.

3.2 Results

Overclaiming in the Math task is defined as a higher contribution amount than the actual percentage of correctly solved squares submitted. Mechanically, each participant contributes one-third *on average*. If participants do not overclaim, claims should also average around 33 percent. As shown in Table 10, over 60 percent of participants overclaim their absolute contribution, with the average claim exceeding 40 percent. This demonstrates that overclaiming is a robust phenomenon that occurs across tasks and subject populations.

We also elicit participant guesses over what percentage of their group's squares they

solved. This helps us understand whether participants have biased recollection of this key objective marker of their contribution or if they simply incorporate other aspects of their performance in their credit claims. Participants overestimate this metric to a similar degree, with guessed contribution exceeding 40 percent, suggesting that some form of self-serving bias in basic recollection of events is at work.

Table 10: Summary Statistics for the Math Tasks

	Delay	Public - Chat	Public - No Chat
Claimed % Contribution	43.85%	40.77%	44.29%
Guess % Contributed	40.43%	41.97%	41.19%
% (Claim > Contributed)	65.7%	58.9%	70.5%
% Overclaim Rank	27%	23.8%	43.1%
% (Guess > Contributed)	49.4%	41%	58.8%

Note: Claimed % Contribution is self-assigned contribution, while *Guess % Contributed* is the guess of their percentage of correctly solved squares. The last three variables represent percentages of the sample who claim to contribute more than they do (*% (Claim > Contributed)*), who claim a higher amount such that their relative rank is higher than they actually achieve (*% Overclaim Rank*), and who guess they submit a higher percentage of correct squares than they actually do (*% (Guess > Contributed)*).

While overclaiming occurs in all treatments, there is heterogeneity in the extent of overclaiming. Table 11 shows the result for a series of linear regressions to examine treatment effects using *Public - Chat* as the omitted category. First, similar to the Riddle task, the measure of overconfidence derived from the science quiz result is not a significant factor for any measure of overclaiming, further establishing it as its own phenomenon, even in more objective tasks. Columns (1) and (2) use the absolute amount overclaimed as the outcome variable. While there is no significant difference between the omitted *Public - Chat* condition and the *Public - No Chat* condition, adding a *Delay* to the contribution elicitation has a marginally significant effect on increasing the amount overclaimed (p -value $\approx 8.9\%$). When considering a binary outcome variable for overclaiming in columns (3) and (4), there is a marginally significant impact of being in *Public - No Chat* compared to *Public - Chat*. Specifically, removing the chat increases the likelihood of overclaiming by 12.7 percentage points (p -value $\approx 6.7\%$).

Turning to columns (5) and (6), we see that overestimating one's relative contribution is 20 percentage points more likely ($p < 1\%$) with no chats. Discussing contributions with teammates de-biases participants more so than just publicly disclosing the contribution claims. Overall, 90 percent of groups used the chat to discuss their contributions. 63 percent of participants come to a consensus and agree on the amount to assign themselves. Not agreeing,

Table 11: Math Treatments Overclaiming Regression

	(1) Δ Overclaim	(2) Δ Overclaim	(3) Binary Overclaim	(4) Binary Overclaim	(5) Rank Overclaim	(6) Rank Overclaim	(7) Guess Overclaim	(8) Guess Overclaim
Public - No Chat	3.480 (3.043)	3.535 (2.999)	0.117 (0.075)	0.127* (0.070)	0.193** (0.077)	0.195*** (0.072)	0.023 (0.079)	0.020 (0.078)
Delay	3.096** (1.817)	2.943 (1.815)	0.0678 (0.048)	0.073 (0.048)	0.033 (0.043)	0.036 (0.041)	-0.020 (0.049)	-0.024 (0.049)
Overconfident		-1.501 (1.730)		-0.009 (0.043)		0.008 (0.038)		-0.002 (0.046)
Gender		-1.488 (1.690)		0.075* (0.043)		0.003 (0.039)		0.029 (0.045)
% Contributed		-0.329*** (0.124)		-0.0133*** (0.003)		-0.021*** (0.004)		-0.005 (0.004)
Constant	7.500*** (1.419)	18.69*** (4.979)	0.589*** (0.038)	0.978*** (0.139)	0.238*** (0.033)	0.948*** (0.150)	0.565*** (0.038)	0.660*** (0.140)
<i>N</i>	496	494	496	494	496	494	496	494
Controls	No	Yes	No	Yes	No	Yes	No	Yes

Note: Standard errors in parentheses clustered at individual level. Controls include demographic information such as major and age. Omitted group is Public-Chat. Dependent variables are as follows: (1) and (2), Δ Overclaim: the difference between self-rated and actual contribution. (3) and (4), Binary Overclaim: binary indicator of if own claim is higher than actual contribution. (5) and (6), Rank Overclaim: binary variable indicating if claimed relatively higher in rank than actually assigned. (7) and (8), Guess Overclaim: binary variable for guessing that they submitted a higher % of correctly answered squares than they actually did. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

either through not discussing in the chat or not reaching a consensus, leads to claiming a higher amount than what the rest of the group assigns that participant (Agree: $\mu = 0.613$, $sd = .343$; Not Agree: $\mu = 6.53$, $sd = 2.39$; $t = 3.816$; $p < 0.01$), which in turn led to overestimating contribution (Agree: $\mu = 5.19$, $sd = 1.60$; Not Agree: $\mu = 11.43$; $sd = 2.63$; $t = 2.143$, $p = 0.016$). The chat logs also reveal that discussions about who contributed most or least happen more often than discussions about exact contribution levels. This suggests that chatting could increase the salience of others’ contributions or aversion to lying, especially about the rank as opposed to numeric claims that do not change implied rank.

3.3 Subjective Evaluations

We also asked participants in the Math task to make subjective qualitative assessments of their contribution with questions adapted from Exley and Kessler (2019) in contrast to our numerical contribution claims.¹⁴ Participants rate their “helpfulness” on a 7-point Likert scale and state how much they agree with the statement “I was helpful” on a 100-point scale for the math task.¹⁵

Table 3.3 presents the subjective helpfulness ratings for the Math task. On average, participants rated themselves as between “Good” and “Very Good” in terms of their helpfulness. Similarly, participants reported 85 percent agreement with the statement “I was helpful.” on average. Curiously, almost 40 percent of those participants who assign themselves *zero* percent contribution still agree with that statement. We find a significant treatment effect on willingness to engage in self-promotion (here defined as saying you’re more helpful). Specifically, we find that those in the *Delay* treatment rate themselves significantly higher in subjective helpfulness on a Likert scale. This suggests that subjective evaluations are more easily influenced by memory-induced biases. Finally, the subject evaluations are not significantly correlated with our measure of overclaiming, but are significantly correlated with overconfidence.

4 Conclusion

Accurate assessment of contributions to group work are important in many settings. Self-reported claims are often used as a proxy for hiring and promotion decisions. We document

¹⁴These questions were also asked in *Delay* sessions of the Riddle task, and results are presented in the Appendix.

¹⁵Results from similar questions asked about the science quiz are in the Appendix.

	(1) Leikert	(2) Leikert	(3) Agree	(4) Agree
Public - No Chat	0.173 (0.174)	0.132 (0.138)	2.429 (2.881)	1.647 (2.291)
Delay	0.181* (0.101)	0.160* (0.089)	2.766 (1.777)	2.384 (1.514)
Claim		0.005** (0.002)		0.075** (0.035)
% Contribute		0.072*** (0.010)		1.296*** (0.205)
Overclaimed		0.084 (0.095)		2.265 (1.572)
Overconfident		0.136* (0.080)		2.825** (1.301)
Constant	4.494*** (0.081)	1.773*** (0.335)	85.24*** (1.541)	36.14*** (7.086)
<i>N</i>	496	496	496	496

Note: Standard errors in parentheses clustered at individual level. Claim is the contribution claim submitted. % Contribute is actual contribution. Overclaimed is if Claim \geq % Contribute. Overconfident is an overconfidence measure from the science quiz. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

consistent overclaiming across different tasks (both a subjective riddle task and an objective Math task) and sample populations (both in the lab and online). Our results suggest that it is a distinct behavioral bias that is not simply stemming from overconfidence. It is thus important to understand the underlying mechanisms and interventions at the self-reporting stage that might reduce overclaiming.

Participants in our study collaborate in groups of three on real effort tasks before making claims their contribution and their teammates'. Over seventy percent of participants ascribe more credit to themselves compared to objective third-party raters. Thirty-three percent of participants assigned contributions in such a way that the implied ranking of who contributed most to least differs from the third-party rankings. Thus, eliciting a ranking of who contributed most to least rather than absolute percentages of contribution can dampen the effects of overclaiming.

In the *Public* treatment, group members know they will be shown each others' contribution claims. Moreover, in the *Public - Chat* treatment, participants are able to first discuss their contributions with their teammates. We find that these treatments significantly reduce overclaiming, especially when participants are able to chat with each other first.

References

- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond**, “Preferences for truth-telling,” *Econometrica*, 2019, *87* (4), 1115–1153.
- Bar-Hillel, Maya, Tom Noah, and Frederick Shane**, “Learning psychology from riddles: The case of stumpers,” *Judgment and Decision Making*, 2018, *13* (1), 112.
- Barber, Brad M. and Terrance Odean**, “Boys will be Boys: Gender, Overconfidence, and Common Stock Investment*,” *The Quarterly Journal of Economics*, 02 2001, *116* (1), 261–292.
- Danz, David, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K. Pun Winichakul**, “Going virtual: A step-by-step guide to taking the in-person experimental lab online,” Technical Report, Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3931028> 2021.
- Exley, Christine L and Judd B Kessler**, “The gender gap in self-promotion,” Technical Report, National Bureau of Economic Research 2019.
- Hügelschäfer, Sabine and Anja Achtziger**, “On confident men and rational women: It’s all on your mind(set),” *Journal of Economic Psychology*, 2014, *41*, 31–44. From Dual Processes to Multiple Selves: Implications for Economic Behavior.
- Niederle, Muriel and Lise Vesterlund**, “Do women shy away from competition? Do men compete too much?,” *The quarterly journal of economics*, 2007, *122* (3), 1067–1101.
- Zimmermann, Florian**, “The Dynamics of Motivated Beliefs,” *American Economic Review*, February 2020, *110* (2), 337–61.

Appendices

A Riddle Task

A.1 Overview of Riddles Used in Main Experiment

Riddle 1: “Bus tickets can be bought one at a time for one dollar each or you can buy a card for five dollars that is good for five rides. Bus drivers can give change. A first time passenger boards the bus, and hands the driver five dollars. The passenger says no word, makes no gesture, and shows no written sign. The bus driver takes her money, and hands her a five-rides card, with one of the five holes punched out. The passenger thanks the driver and takes a seat. How did the driver know that the passenger wanted the five-ride card rather than a single ride?”

Riddle 2: “A black riderless horse was standing in the middle of a black asphalt road. The streetlights were not on, and the moon was not out. A car was travelling towards the horse at full speed, its headlights off. Yet, the driver managed to see the horse in time to brake and avoid crashing into it. How is that possible?”

Riddle 3: “In a Bangladesh market, a small potato bag costs 5 taka, a medium potato bag costs 7 taka, and a large potato bag costs 9 taka. Yet, a single potato in that market costs 10 taka. How is that possible?”

Table A1: Summary Statistic for Riddles Used in Main Experiment.

	Riddle 1	Riddle 2	Riddle 3
N	401	206	169
N – Baseline	89	122	80
N – Delay	111	84	0
N – Public	201	0	89
Correct Individual Answers	48.88%	91.26%	79.88%
Correct Group Answers	60.69%	97.54%	95.85%

Note: ‘Correct Individual Answer’ details the percent of subjects who submitted a correct answer. ‘Correct Group Answer’ details the percent of groups where at least one subject responded correctly.

A.2 MTurk Rating Session

After the riddle task, we ran a second session where other mTurk workers rated contribution of group members. Subjects in the rating session were presented with the riddle, the transcript of their chat, and the answers submitted.¹⁶

Each subject in the rating session saw between three and five chats, and each chat was rated by between three and five people. Subjects in the rating sessions were paid \$1.50 as a HIT fee and an addition \$0.50 for each group for which they were within 20 pp of the answer of the average rating.

For a subset of early samples, we employ two separate groups of mTurk workers as raters. Each contribution was therefore rated six times, instead of the usual three. One group rates contributions using the default set-up employed across all ratings, tasks, and treatments: they see chat-logs with gender-blind nicknames (ie., replacing the names shown in the chat transcript by ‘Player 1’, ‘Player 2’, ‘Player 3’, coloring of chat participants name were in gender-neutral colors, yellow, gray, and orange). The other group sees the same nicknames as occur in the chats, i.e., revealing gendered information.

We find no evidence that revealing gender has any impact on contribution rating, for neither male nor female raters. Figure A1 displays the average rating for each subject in the subsample under the revealed-gender condition plotted against the average rating under the unknown-gender condition. We see a tight fit of the two regression lines against the red 45-degree line: in orange, we plot the regression of average score assigned on the known-gender condition against the average score in the unknown-gender condition for women; in blue, we plot the same regression for men. The two regression lines have virtually identical slopes and intercepts. The scatterplot reveals no relevant differences in non-linear effects or outliers across the two genders.

A.3 Chat Analysis

To better understand how contribution was evaluated, we examined the impact of a variety of variables from the riddle answer discussion on the average rating of contribution. The number of words and characters typed, along with the amount of distinct messages sent, did not significantly impact ratings by either the self or by the third-party raters. There did appear to be a first mover advantage, with sending the first message (which was usually uninformative, such as “what do you think?”) being associated with over a 5 percentage

¹⁶Raters would only see groups who solved the same riddle.

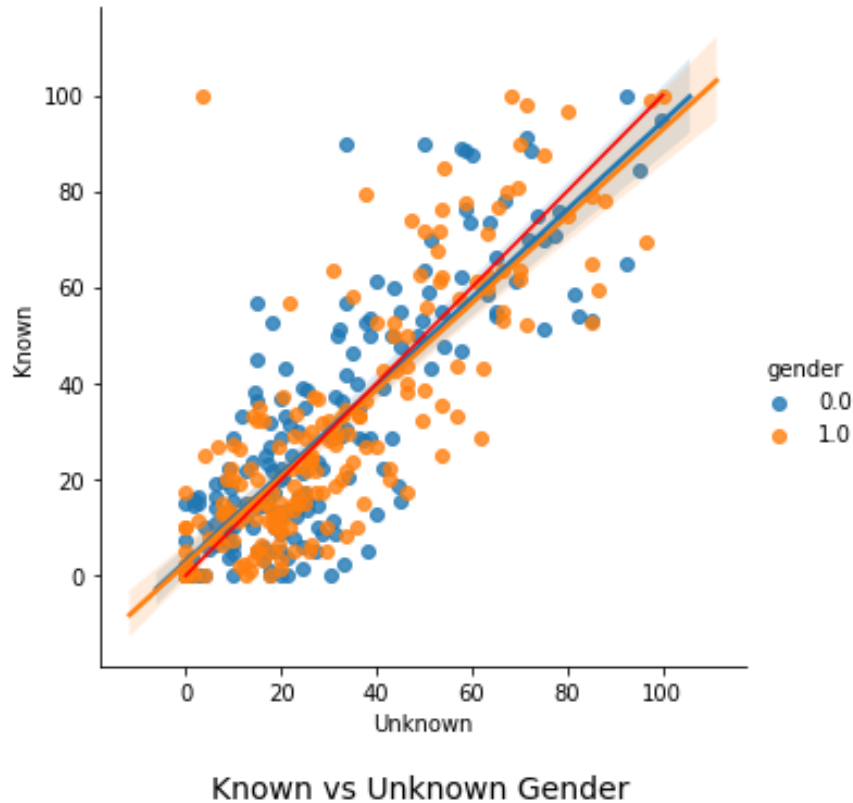


Figure A1: Known versus Unknown Gender – Average Ratings.

Note: The average score in the unknown-gender condition is plotted on the x-axis and the average score in the known-gender condition on the y-axis. Women are in orange, men in blue.

point increase for both self- and third-party- ratings. The impact was larger for third-party raters, but not significantly so. On the other hand, having a back-and-forth conversation (defined as sending two messages with someone else messaging in between) was associated with significantly lower ratings.

B Math Task

B.1 Performance Statistics

Table B1 summarizes the actions of participants in the Math task, pooling across treatment groups. We see a small and marginally significant male advantage in the task, as men complete on average 33.8 percent of all squares compared to the 32.8 percent completed by women (versus the mechanical mean of 33.3 percent for the whole sample) (p -value of

Table A2: Text Analysis of Riddle Chat Data on Rating

	Self	Third Party
First	5.430** (2.401)	6.034* (1.828)
Words	0.465 (0.475)	0.101 (0.363)
Characters	-0.0589 (0.0916)	0.0313 (0.0700)
Distinct	1.122 (0.987)	-0.212 (0.755)
Back and Forth	-2.582** (1.183)	-1.865** (0.905)
Constant	44.86* (1.821)	30.00* (1.385)

Note: Standard errors in parentheses clustered at the individual level. Outcome variable is the average credit claim assigned based off different chat characteristics while discussing the riddles answers. Controls for treatment.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

difference $\approx 9.9\%$). Men spend slightly fewer seconds active than women (p -value $< 1\%$), supply slightly more correct squares (p -value $< 1\%$), and use the skip voting feature more than women (p -value $< 1\%$). Women and men guess incorrectly equally as often and are skipped similarly often.

Table B1: Performance Statistics by Gender for the Group Math Task

	Female	Male	Difference	se
% Contribute	32.808	33.790	-0.982*	0.594
Active Seconds	100.862	93.678	7.184	1.636***
# Correct	7.157	7.923	-0.765***	0.188
# Incorrect	0.931	0.816	0.114	0.091
Used Skip	1.348	2.277	-0.929***	0.303
Was Skipped	0.275	0.209	0.066	0.098
N	261	234	495	

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C Subjective Evaluations

These questions were adapted from Exley and Kessler (2019). In Exley and Kessler (2019), women rate themselves subjectively lower on the task at hand, which in their case was an ASVAB test. To test whether we replicate this behavior using our 20-question science quiz, a similar task to that in the original study, we also ask participants about their subjective evaluation of their performance on this quiz on similar scales phrased as “I performed well,” therefore using the original wording from the Exley and Kessler (2019) study.

In Table C1, we summarize Exley and Kessler-style results for the mTurk sample that performed the Riddle task in the *Delay* treatment. In column (1), we find no significant differences between quiz performance of men and women. In columns (2) and (3), we assess the confidence in having performed well on the science quiz, with the outcome variable of the guessed number of correctly solved questions. Here we see no significant differences in beliefs between men and women. In columns (4) through (7) we assess whether the more subjective evaluation of whether the participant performed well – via the Likert scale (4, 5) or via the 100-point agreement scale (6, 7) – leads to a difference in self-evaluations between men and women. We find only limited evidence for this, with only one coefficient significantly different from zero across the four models.

Table C1: Riddle Task Replication

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Score	Belief	Belief	Likert	Likert	Agree	Agree
Male	-0.0127	1.199	1.203	0.512***	0.809	10.96	13.27
	(0.260)	(0.678)	(0.690)	(0.0383)	(0.751)	(4.642)	(14.17)
Quiz Score			0.327				
			(0.129)				
Belief					0.242**		4.854***
					(0.038)		(0.354)
MaleXBelief					-0.0530		-0.734
					(0.081)		(1.149)
Constant	9.094***	9.882***	6.905*	2.988***	0.596	48.25***	0.280
	(0.219)	(0.537)	(1.708)	(0.023)	(0.373)	(3.514)	(2.514)
<i>N</i>	171	171	171	171	171	171	171

Note: Standard errors in parentheses clustered at the individual level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

In the PEEL sample that performed the Math task, however, men scored a small, but significant, amount higher than women on the science quiz, see column (1) in Table C2. In

columns (2) and (3), we further find that men guess having a higher score on the quiz. When considering the subjective rankings (columns (4) and (6)) without covariates, we find that men agree significantly more than women with the statement “I performed well on the task,” both when measured using a Likert scale and when using 100-points ranking scale. However, once we control for beliefs about performance, there are no significant gender differences and no significant interaction between beliefs and subjective rating (columns (5) and (7)). Therefore it appears that, in our sample, the subjective ratings are capturing overconfidence within the science test.

Table C2: Math Task Replication

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Score	Belief	Belief	Likert	Likert	Agree	Agree
Male	0.757*** (0.226)	1.423*** (0.292)	1.019*** (0.268)	0.381*** (0.101)	-0.210 (0.413)	4.706*** (1.806)	-0.918 (7.631)
Quiz Score			0.528*** (0.064)				
Beliefs					0.203*** (0.018)		3.978*** (0.353)
MaleXBelief					0.0224 (0.030)		-0.002 (0.560)
Constant	11.86*** (0.217)	11.91*** (0.291)	5.660*** (0.840)	3.484*** (0.099)	1.073*** (0.235)	64.14*** (1.811)	16.78*** (4.592)
<i>N</i>	495	494	494	494	494	494	494

Note: Standard errors in parentheses clustered at the individual level. Controls for treatments.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Unlike Exley and Kessler (2019), we do not find a difference in willingness to engage in self-promotion (here defined as rating oneself as more helpful) by gender.

Table C3: OLS Estimates of Subjective Measures of Contribution

	(1)	(2)	(3)	(4)
	Likert	Likert	Agree	Agree
Male	0.128 (0.091)	0.065 (0.082)	2.181 (1.511)	1.049 (1.331)
Public - No Chat	0.190 (0.173)	0.158 (0.140)	2.720 (2.869)	2.169 (2.324)
Delay	0.184* (0.100)	0.166* (0.088)	2.834 (1.779)	2.519* (1.520)
Claim		0.006*** (0.002)		0.101*** (0.031)
% Contribute		0.070*** (0.010)		1.254*** (0.199)
Constant	4.429*** (0.093)	1.879*** (0.309)	84.13*** (1.767)	38.85*** (6.641)
<i>N</i>	495	495	495	495

Note: Standard errors in parentheses clustered at individual level. Omitted group is Public - Chat. Claim is how much contribution the participants claimed in the quantified section, while % Contribute is what percentage of correctly solved squares they submitted personally.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

D Experiment Screenshots

D.1 Riddle Experiment – Baseline Screenshots

Tell us about yourself.

Before we start the study, please tell us a little bit about yourself so that we can determine your eligibility to proceed. If you are not selected to proceed, you will be paid the \$0.50 payment.

What is your favorite color?

How old are you (in years)?

What sex do you identify with?

- Male
- Female

What is your favorite music genre?

Next

Figure D1: Demographics

Welcome!

Thank you for your interest in our study! This HIT will take about 30 minutes. You will receive a bonus of \$0.50 for completing this survey. You can earn bonus payments depending on your answers: you can earn **up to \$9.30** and most participants earn at least **\$5.50**.

This HIT involves group work with other MTurk workers.

- Please be respectful of their time and effort.
- Only take this HIT if you can commit to working on it continuously with your full attention.
- If you cannot do so, please abort this HIT now. Thank you.

This study has three parts.

1. You will answer a science quiz.
2. Then, you will work with two other mTurk workers on solving a riddle.
3. Finally, you will complete a survey on your own.

You can earn bonus payments **in each of the three parts** of the survey. You will not be paid any **bonus** unless you complete the survey. Please read all instructions carefully, as this will give you the best chance at maximizing your earnings.

All instructions in this survey are completely honest and truthful. The Stanford University Economics Department and Professor Doug Bernheim, principal investigator of this IRB protocol and Department Chair, guarantee this. In particular, when we say that the computer will draw a random number, we will use real cryptography-grade randomization devices.

Next

Figure D2: Introduction

Quiz Instructions

You will answer 20 quiz questions. Please refrain from using the internet to find answers. You have 15 seconds per question, so looking up answers will be slower than just thinking about your best guess.

For each question, there are four possible answers and one answer is correct. After you have submitted your answer (by clicking on the "Next" button), you **cannot** go back to change your answer.

We will pay you \$0.05 for **each** question you answer correctly!

Next

Figure D3: Quiz Instructions

Question 1 of 20

Time left to complete this page: 0:11

Bone is composed primarily of which inorganic material?

- Calcium
- Phosphorus
- Collagen
- Potassium

Next

Figure D4: Sample Quiz Question

Next, you will be matched with two other MTurk workers.

On the next screen, you will be matched with two other MTurk workers.

- Be respectful of their time and effort and only continue this HIT if you can work on it continuously.
- You still need at least 25 minutes for this HIT.
- You only receive a bonus if you complete this HIT.
- Please be patient if other workers are slower or if it takes some time until you are matched with other workers.

Thank you.

Next

Figure D5: Matching Page

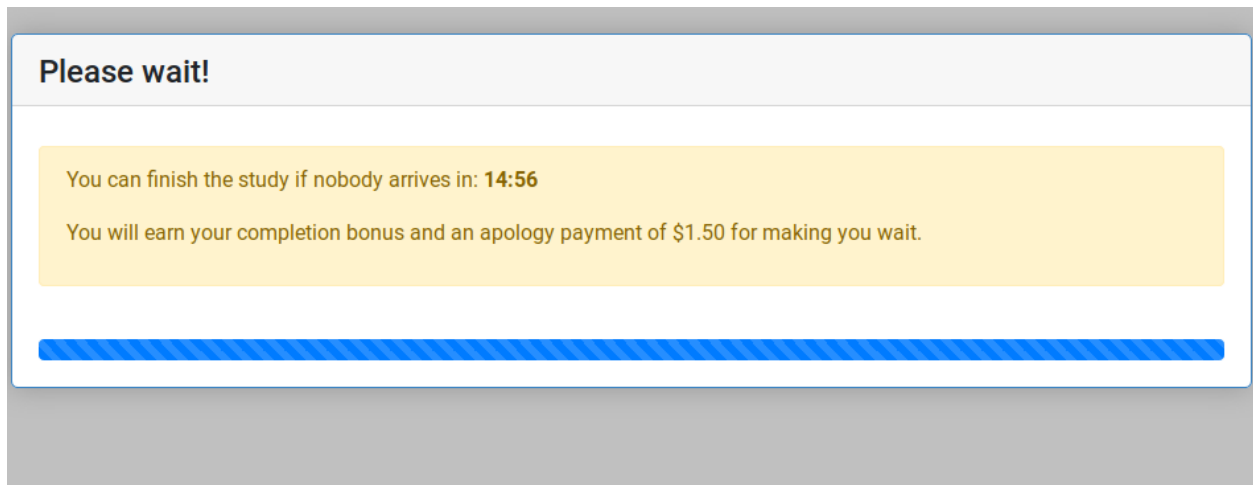


Figure D6: Matching Wait Page

Meet Your Team-Mates!

You have been matched with two other players:

Player 1 in your team is:

Jennifer.

Jennifer is 22 years old, likes the color Pink, and listens to Pop.

Player 2 in your team is:

Mary.

Mary is 46 years old, likes the color yellow, and listens to Classical.

Together you will try to solve a riddle. You will be able to discuss the riddle with your team-mates in a chat-box. Please make use of this to brainstorm about ideas for solving the riddle.

Each of you will individually submit an answer. The computer will randomly select one of the three answers. If the randomly chosen answer is correct, each of you will get a **\$2.00 bonus!**

Figure D7: Introducing Team Mates

Riddle

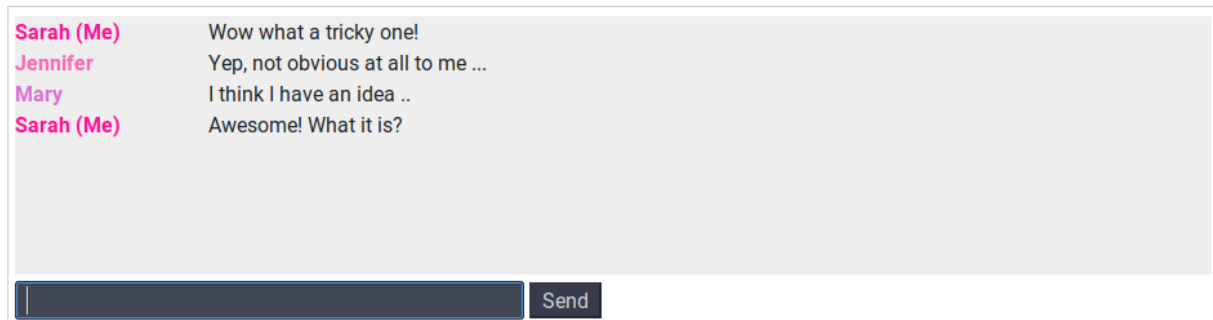
Time left to complete this page: 14:51

The riddle is:

A black horse was standing in the middle of a black asphalt road. The streetlights were not on, and the moon was not out. A car was travelling towards the horse at full speed, its headlights off. Yet, the driver managed to see the horse in time to brake and avoid crashing into it. How is that possible?

Discuss the riddle in the chat and when you feel confident in an answer, enter it below.

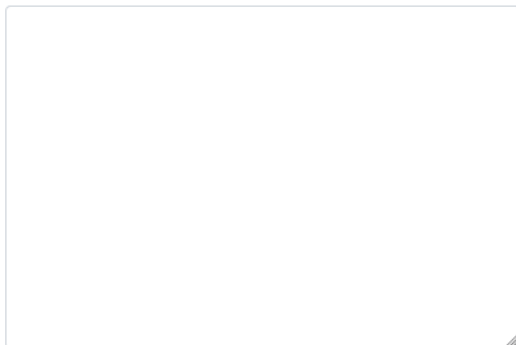
Figure D8: Riddle



A screenshot of a chat interface. The chat area is a light gray rectangle containing four messages. Each message has a name in pink on the left and the text of the message on the right. Below the chat area is a dark gray input field with a white cursor and a dark gray 'Send' button to its right.

Sarah (Me)	Wow what a tricky one!
Jennifer	Yep, not obvious at all to me ...
Mary	I think I have an idea ..
Sarah (Me)	Awesome! What it is?

What is the solution to this riddle?



A large, empty rectangular box with a thin gray border, intended for the user to write their answer to the riddle.

Figure D9: Chat (on same page as riddle)

How much did you contribute?

Please think about how much **you** contributed to solving the riddle. Use the slider, from 0% contribution to 100% contribution. Note that the starting position of the slider is randomly generated.

You can earn a bonus payment of \$1.00. We have designed the payment scheme such that it is in your best interest to simply think carefully about your contribution and then tell your best estimate. Your answer does not affect your team-mates' bonus payments.

For details on the payment rule, click on the 'Details' tab below. We will have a group of mTurk workers read your chat messages. We will anonymize the chat names by replacing them by 'Player 1', 'Player 2', and 'Player 3'. These workers will grade you and your team-mates' contributions to solving the riddle and assign each of you a score between 0% and 100%.

How much did you contribute to solving the riddle?



Sarah (Me)	Wow what a tricky one!
Jennifer	Yep, not obvious at all to me ...
Mary	I think I have an idea ..
Sarah (Me)	Awesome! What it is?

[Details of the Payment Rule](#)

Next

Figure D10: What was your contribution?

Quiz Instructions

You will answer 20 quiz questions. Please refrain from using the internet to find answers. You have 15 seconds per question, so looking up answers will be slower than just thinking about your best guess.

For each question, there are four possible answers and one answer is correct. After you have submitted your answer (by clicking on the "Next" button), you **cannot** go back to change your answer.

We will pay you \$0.05 for **each** question you answer correctly!

Next

Figure D11: Details of Payment Rule for Contribution Elicitation

How much did your team-mates contribute?

You assigned yourself a contribution of 59%. How much did each of your team-mates contribute to solving the riddle?

You can earn a bonus of \$1.00 depending on your answer. It is in your best interest to answer truthfully and accurately. For details on the payment rule, click on the 'Details' bar below. The two sliders add up to 41% automatically, for a total of 100%, and their starting positions are randomly chosen.

Your answers **do not** affect your team-mates' payoff.

First, think about **Jennifer**.

How much did this player contribute to solving the riddle?



Next, think about **Mary**.

How much did this player contribute to solving the riddle?



Sarah (Me)

Wow what a tricky one!

Jennifer

Yep, not obvious at all to me ...

Mary

I think I have an idea ..

Sarah (Me)

Awesome! What it is?

Next

[Details of the Payment Rule](#)

Figure D12: What was the contribution of your team mates?

Extra Bonus: How well did you do in the quiz?

Please think back to the science quiz you took at the start of the experiment.

How many of the 20 quiz questions do you think you answered correctly?

Please think carefully! The closer your answer is to the truth, the more you will earn: if this question is selected for payment, you can get an extra bonus of up to **\$2.00!**

[Details of the Payment Rule](#)

Next

Figure D13: How well did you do on the quiz? (Overconfidence elicitation)

Details of the Payment Rule

The computer will check your quiz score and then compute the difference between your quiz score and your answer to the question here.

Then, the computer will subtract the square of this difference divided by 25 from \$2.00.

In numbers:

Your bonus = $\$2.00 - 1/25 * (\text{Your Answer} - \text{Your Actual Score})^2$

This may sound complicated but rewards you for giving your best possible guess!

Figure D14: Payment Rule Quiz Confidence

D.2 Math Experiment – Baseline Screenshots

Many of the pages for the Math Experiment are analogous to the Riddle experiment. We highlight the relevant differences between the two experiments by displaying the different screens below:

Welcome!

The session today has four parts:

1. You will answer a science quiz.
2. Then, you will try to solve as many adding-up tasks as possible in 3 minutes
3. Then, you will work with two other participants to solve as many adding-up tasks together as you can in 5 minutes.
4. Finally, you will fill out a short survey on your own.

You can earn bonus payments **in each of the parts** of the study. Please read all instructions carefully, as this will give you the best chance at maximizing your earnings.

All instructions in this study are completely accurate and truthful. The University of Pittsburgh Economics Department and Stephanie Wang, the principal investigator, guarantee this. In particular, when we say that the computer will draw a random number, we will use real cryptography-grade randomization devices.

If you have any questions at this time, please send the researcher a chat message on Zoom.

Figure D15: Welcome Screen – Math

Instructions - Adding Up Task

Your task is to add up the nine numbers inside a three-by-three square of numbers. You will have 180 seconds to solve as many of these squares as possible. For each square you solve correctly, we will add \$0.20 to your final pay.

See an example square below:

3	-3	3
-1	-1	1
-1	2	1

The correct answer in this example would have been **4**.

The solution is found by calculating:

$$3 - 3 + 3 - 1 - 1 + 1 - 1 + 2 + 1 = 4$$

If you have any questions please send the researcher a message over Zoom.

Figure D16: Math Task Instructions for Practice Round

Meet Your Team-Mates!

You have been matched with two other players!

Player 1 in your team is:

Jennifer

Jennifer is 27 years old, likes the color blue, and listens to rock.

Player 2 in your team is:

Robert

Robert is 18 years old, likes the color red, and listens to pop.

Figure D17: Meeting Team Members pt.1 – Math

Group Adding-Up Task

Your task is the same as in the last part of the study: add up the nine numbers inside the square, to earn \$0.20 bonus for each correctly solved square.

However, this time you will work as a team. For every correct answer submitted by your team, you will get a bonus payment of \$0.20. This means that **every team-member** in the team will get \$0.20 for every correct answer submitted by any of the team-members.

For each square, one participant will be the **active** player and can submit an answer. The other two players will be **reserve** players for that square. After the active player submits an answer, one of the reserve players will become the active player and gets to solve a **new** square. The reserve players can vote to skip the active player by pressing a "skip" button. If both reserve players vote to skip, one of them becomes the active player and can submit an answer for **that** square. After the new active player submits their answer for that square, the next reserve player will become the active player and will solve a new square.

After the group adding-up task, you will fill out a survey. After the survey, you will be given information about how to receive your payment. Specifically, you will be given a link to a secure form where you will provide your Venmo information and will act as your digital receipt. Please ensure that your information is correct, as after the study we will send your payment to the exact Venmo information you provide.

Please send me a chat message over Zoom if you have any questions.

Figure D18: Meeting Team Members pt.2 – Math

Time left for this quiz: 2:58

Please add up the nine numbers in the square:

1	-1	2
3	-1	2
2	1	1

Once you feel confident in an answer, enter it below and click on the green "Answer" button.

Figure D19: Active Player before submitting an answer

Time left for this quiz: 2:13

Your solution was **INCORRECT**.

Please add up the nine numbers in the square:

-1	1	-2
2	-2	1
2	-3	-3

Once you feel confident in an answer, enter it below and click on the green "Answer" button.

Figure D20: Active Player before submitting an incorrect answer

Time left for this quiz: 2:46

Your solution was **CORRECT**.

Please add up the nine numbers in the square:

1	3	-1
1	3	-1
-2	-1	2

Once you feel confident in an answer, enter it below and click on the green "Answer" button.

Figure D21: Active Player before submitting a correct answer

Time left for this quiz: 4:58

You are a reserve player

Sarah is working on this square.

-3	2	3
1	1	-3
-1	2	-2

If you want to vote for ending **Sarah's** turn, click below.

SKIP ACTIVE PLAYER?

Figure D22: Reserve Player