Pre-Analysis Plan

For "Gender Differences in Self-Promotion and Career Advice" by Rigissa Megalokonomou, Juliana Silva Goncalves and Roel van Veldhuizen

A. Description of the Experiment

We conduct a three-part experiment. In session 1, participants in the role of workers work on a math and science quiz. Upon completing the quiz, they indicate their confidence in several ways, forming a worker "portfolio". In session 2, a different set of participants in the role of advisers sees the portfolio of five workers and advises each worker whether they should go for an advanced (A) or a basic (B) version of the task. In session 3, each worker receives the advice from one adviser and decides which of the two versions of the task to pursue.

We consider three between-subject treatments that vary the information the adviser receives in session 2. In the "baseline" treatment, the adviser receives the full portfolio of worker confidence measures from session 1. In the "nudge" treatment, advisers also receive a brief message informing them about the gender gap in self-assessment. In the "subjective" treatment, the advisers instead only receive the worker's verbal self-evaluation (a qualitative and subjective self-assessment). In addition, participants in the baseline treatment (but not the other treatments) are also asked to evaluate the same five workers a second time while also having access to those workers' true performances on the task. We refer to this second set of advice as the (within-subject) "objective" treatment.

B. Main Outcomes

1. Worker Confidence (Session 1)
    a. Guessed number of correct answers
    b. I performed well on the quiz (0-100)
    c. I would apply for a job that would require me to perform well on the quiz (0-100)
    d. I would succeed in a job requiring me to perform well on the quiz (0-100)
    e. Relative performance rank (compared to 100 other participants)
    f. Subjective performance Likert scale (1-terrible to 5-great).
    g. Standardized sum of the previous measures
    h. Verbal self-evaluation
2. Advice (Session 2)
    a. Binary advice (A or B)
    b. Five-point scale advice (Definitely A to Definitely B)

C. Secondary Outcomes

3. Worker Task Choice (Session 3)
4. Other measures of advice (Session 2)
    a. Performance target for workers
    b. Confidence in binary advice

D. Main Explanatory Variable

5. Worker Gender: dummy variable comparing males to females.

E. Other Variables

6. Workers (session 1): general confidence/risk aversion measures, score, demographics
7. Advisors: gender, other demographics, confidence/risk aversion
8. Advisor beliefs about gender differences:
    a. Are men or women better at the task?
    b. Are men or women more confident in their self-assessment?
9. Worker score in session 3.

Demographics include household income, income satisfaction, age, occupation, job title, and education.

F. Main analysis

Our experiment aims to test the following hypotheses:

1. Gender Gap in Self-Promotion: Female workers are less confident than male workers.
    a. Main tests: one-sided t-test for each of the quantitative worker confidence variables listed in part B1 above (a-g).
2. Gender Gap in Advice: Female workers receive less ambitious advice than male workers in the baseline treatment.
    a. Main tests: linear regression of the (a) binary advice measure and (b) Likert scale advice measure on a dummy for the worker's gender. We use a one-sided test on the gender dummy.
3. Treatment Differences: The gender gap in advice is significantly smaller in the objective, nudge and subjective treatments than in the baseline.
    a. Main tests: a linear regression of advice on gender, treatment and a gender*treatment interaction. The key test is a one-sided test on the significance of the gender*treatment interaction. We do three regressions comparing each of nudge, subjective and objective treatment to the baseline.

We will use robust standard errors clustered at the referee level for the regression analysis (F2- F3). In addition to testing these three hypotheses, we will also investigate the drivers of the gender gap in advice:

4. Drivers of the Gender Gap in Advice: Is the gender gap in advice (F2) driven by gender differences in self-promotion or by a gender bias among advisers?
    a. Main tests: linear regression of the (a) binary advice measure and (b) Likert scale advice measure on (i) a dummy for the worker's gender and (ii) the worker's guessed score (B1a). We use a one-sided test for both variables (less ambitious advice for women, positive effect of the guessed score).

G. Additional Analysis

Here is an overview of additional tests we intend to include in our paper to shed further light on our results, but that do not directly address our main hypotheses. We start with some further analysis of worker confidence.

5. Is there a gender difference in performance?
    a. Two-sided t-test for a gender difference in worker score (session 1).
    b. Two-sided t-test for a gender difference in worker score (session 3), after receiving advice.
6. Are any gender differences in confidence significant after controlling for performance and demographics?

    a. Regression of confidence on the worker gender dummy (1-female, 0-male) while controlling for (a) worker score from session 1 and (b) also controlling for worker demographics. The relevant test is a one-sided test on the regression coefficient on the female dummy (as per main analysis F1).
7. Are there any gender differences in the subjective self-evaluations?
    a. For this purpose, we intend to explore the use of AI methods to quantify the expressed levels of confidence.

Next we plan to present robustness checks for our analysis on gender differences in advice, as well as some additional outcomes.

8. Is the gender difference in advice F2 robust to only looking at the first worker evaluated?
    a. Replication of main analysis F2 while including only the first worker evaluated.
9. Do gender differences and treatment differences in advice (F2 and F3) also appear in our secondary measures (C4)?
    a. Replication of main analysis F2 and F3 for the performance target C4a.
    b. Replication of main analysis F2 and F3 using the confidence in advice variable (C4b). For this purpose, we combine the binary variable with the confidence variable in the following way: we will code advice given with confidence 1-3 as "no advice", confidence 4-8 as "probably A/B" and 9+ as "definitely A/B" and encode this as a five-point scale from 1-definitely B to 5-definitely A, so as to approximate the distribution of responses on the Likert scale (based on pilot data).
10. Does the adviser's confidence (C4b) depend on the worker's gender and advice sent?
    a. Two-sided t-test investigating whether confidence differs significantly by (i) worker gender and (ii) binary advice given (A or B). In addition, two two-sided t-tests studying whether confidence differs significantly by worker gender for a given piece of advice (A or B).
    b. A linear regression of adviser confidence on either (i) worker gender, (ii) binary advice (A or B) or (iii) gender within the sample of a given type of advice (A or B), in all cases while controlling for the worker's guessed score.
11. What are the main drivers of advice?
    a. Linear regressions of advice (B2a and B2b) in the baseline on (i) all quantitative measures of confidence available to advisers (B1a-e) and (ii) referee demographics.
    b. We also intend to explore whether the effect of confidence is non-linear by regressing advice on guessed score fixed effects.
12. What is the quality of advice and how does it differ by treatment and gender?
    a. We create a binary indicator for optimal advice based on whether advice, if followed, would maximize the worker's expected earnings based on their actual score from session 1. We then regress this measure on dummies for the respective treatments. We also replicate this analysis while including a gender dummy and interactions of this dummy with the treatment dummies.
13. Do we observe heterogeneous effects based on adviser bias?
    a. Replicating main analysis F2 while splitting the sample based on advisor bias E8.
    b. We split the sample in three groups: no difference, men at least slightly better/more confident and women at least slightly better/more confident. We only analyze a given group if at least 5% of our sample are in this group.
    c. We look at each variable E8a and E8b separately.
14. Do any of our treatments reduce any gender bias that exists in variables E8a and E8b?
    a. Regression of each of these variables on dummies for the various treatments. We use a two-sided t-test on each of the treatment dummy coefficients.

b. We will only do this test if there is a significant bias in the baseline (one-sided t-test, women being seen as worse/less confident).

We also consider several tests of behavior in Session 3. These are of lesser importance because we (a) expect nearly all workers to follow advice and (b) we have less power due to the smaller sample size per treatment.

15. Do female workers choose less ambitious jobs in session 3?
    a. One-sided t-test for worker choice in session 3 by gender.
16. An analysis of the efficiency consequences of advice.
    a. We create a binary indicator of the optimal choice for whether a worker maximized their earnings based on session 1 performance (ex ante). We then regress this measure on dummies for the respective treatments. We also replicate this analysis while including a gender dummy and interactions of this dummy with the treatment dummies.
    b. We also perform a similar analysis on whether the advice sent by advisers is optimal in that (if followed) it maximizes the expected earnings given session 3 performance.
17. Are gender differences in worker choice in session 3 driven by advice and do they differ by treatment?
    a. We split the sample into two groups depending on whether workers receive the binary or Likert-scale advice.
    b. For the binary variable, we regress choice on a gender dummy and the binary advice variable. We also instead consider a more ordinal scale that takes into account adviser confidence. Specifically, we will code advice given with confidence 1-3 as “no advice”, confidence 4-8 as “probably A/B” and 9+ as “definitely A/B” and encode this as a five-point scale from 1-definitely B to 5-definitely A.
    c. For the Likert scale variable, we regress choice on a gender dummy and the advice received, with 1-definitely B and 5-definitely A.
    d. We also test whether the gender gap in worker choices differs by treatment by regressing worker choice on a gender dummy, treatment dummies and gender*treatment interactions.

Power Calculations

Summary

We conducted power calculations to check our power to detect (a) a gender gap in worker confidence (self-promotions), (b) a gender gap in advice and (c) treatment differences in the gender gap in advice. Based on these power calculations, we aim to collect data from 350 workers and 1750 advisers (700 in the baseline, and 525 in the Nudge and Subjective treatments). With the intended sample size, we will have a power greater than 0.90 to detect (a) gender differences in worker confidence in line with the pilot and previous literature, (b) a gender gap in advice of half the size of the gap implied by our pilots and (c) treatment effects that eliminate 70% or more of the baseline gender gap in advice. In the remainder of this document, we will present the power calculations in greater detail.

H1: Gender Differences in Self-Promotion

Our main interest in this experiment is in testing three sets of hypotheses. Our first hypothesis relates to the gender gap in self-confidence (self-promotion) among workers. We can do power calculations using the confidence data from three pilot sessions (N=301 participants). Table 1 presents the pilot results for our quantitative confidence variables (B1a-g). The gender gap was significant at the 1% level for all variables, and the point estimates for the first five variables are consistent with Exley and Kessler (2022) albeit slightly smaller.[1] We use the averages and standard deviations for men and women to compute the power associated with a one-sided t-test testing whether men are more confident than women. The results suggests that a sample size of 210 workers (105 men and 105 women) is sufficient to achieve a power of 0.80 or greater for all six of our confidence variables, as well as when summing over all variables. A sample size of 290 workers would be sufficient for achieving a power of 0.90.

[1] The gender gaps in Exley and Kessler (2022) for the first five variables are 2.28 (p. 1359), 0.67, 13.83, 17.28 and 16.12 (Table II). They do not elicit the rank guess variable.

Table 1: Power Calculations for the Gender Gap in Self-Promotions

	Score Guess	Perform Well Likert Scale	Perform Well 0-100	Apply Job	Perf Job	Rank Guess	Sum
Men	10.14 (3.95)	3.06 (1.02)	44.82 (22.4)	38.63 (28.1)	39.96 (27.3)	54.44 (24.3)	0.29 (0.98)
Women	8.81 (3.75)	2.64 (1.02)	35.89 (22.5)	25.95 (25.1)	27.66 (26.1)	62.67 (22.8)	-0.29 (0.94)
Diff	1.33 (3.91)	0.42 (1.04)	8.93 (22.9)	12.68 (27.3)	12.30 (27.3)	-8.23 (23.9)	0.58 (1)
Samp (0.80)	210	150	158	112	118	206	70
Samp (0.90)	290	206	220	154	164	284	96

Notes: means (standard deviations). "Score Guess" asks participants to guess the number of questions they solved correctly. "Perform Well" elicits subjective performance on a five-point Likert scale and a 0-100 scale respectively. The next two questions ask, on a 0-100 scale, whether they would apply for a job that requires them to perform well on the quiz or be successful in such a job respectively. The rank guess variable asks participants to guess their rank compared to 100 other participants (1-best, 101-worst). The final variable takes the sum of the standardized first six variables (reverse coding the rank guess variable) and then standardizes this sum. Required sample sizes ("Samp") are based on "power twomeans mu_men mu_women, sd1(sd_men) sd2(sd_women) onesided", where the four parameters are the ones presented in the first two rows of the table (means and sd). We present results for both power calculation with a power of 0.80 ("Samp (0.80)") and for a power of 0.90.

H2: Gender Differences in Advice

Our second hypothesis relates to the gender gap in advice. We conducted pilot sessions with a total of 200 advisers who each evaluated 5 worker portfolios for a total of 1000 adviser-level observations. The results revealed that (a) the various confidence measures were highly correlated with each other ($r>0.6$ in all cases) and (b) advisers relied heavily on the confidence variables in determining their advice. For example, a regression of binary advice on the worker's guessed quiz score yielded a coefficient of 0.067 (se: 0.003, $p<0.0001$). The resulting gender gap in advice received was also significant at 5.6pp ($p=0.015$, one-sided t-test). This gap might understate the true expected gender gap in the main sessions, however, because the sample of workers the advisers evaluated had a lower gender gap in confidence (0.42 points on the score guess variable) than the one we saw when combining worker data from all pilots (1.33 points).

Hence, we compute power in the following way. First, we randomly pick a sample of workers from the pilot sessions with replacement. We then simulate advice data using the pilot coefficient estimates of a regression of advice on workers' guessed quiz score. We also add noise to the simulated advice to the point that the guessed quiz score explains the same fraction of the total variance as in the pilot. We ensure that the Likert scale advice has five values by winsorizing at 1 and 5 and rounding to the nearest integer. Similarly, we binarize the binary advice measure by converting, for example, a predicted advice of 0.75 into either advice A (with 75% chance) or advice B (with 25% chance). Power is then equal to the fraction of simulated samples in which women receive less ambitious advice with $p<0.05$ (one-sided).

Table 2 presents the results. The power calculations predict an approximately 8.6pp gender gap in the binary measure and a 0.25-point gap in the Likert scale measure. Noting that each adviser evaluates five workers, a sample size of 120-150 advisers would already yield power above the 0.80 threshold, where power is slightly higher for the Likert scale measure. Note that this assumes that each piece of advice is an independent observation, which seems reasonable in the present context (and given that the data were

simulated from data in an identical environment). Note also that we conservatively assume that advisers are not intrinsically biased against women; adding such a bias would raise the predicted gender gap and thereby further increase our power.

Table 2: Power Calculations for a Gender Difference in Advice

		Binary Advice		Likert Advice	
#Evaluations	#Advisers	Gender Gap	Power	Gender Gap	Power
600	120	-0.085	0.724	-0.249	0.816
800	150	-0.088	0.876	-0.263	0.930
1000	200	-0.087	0.929	-0.254	0.964
2000	400	-0.086	0.993	-0.251	1

Notes: The table presents results from four sets of simulations that vary the sample size from 600 to 2000 portfolio evaluations (evenly split across the two genders). Binary advice is a dummy variable for whether advice A is given, and Likert advice is a five-point scale ranging from "definitely B" to "definitely A". Each row is based on 1000 simulated samples. "Gender Gap" is the difference between men and women on the relevant outcome averaged across all simulation samples. Power is the fraction of simulated samples in which an OLS regression of the outcome on gender produced a significant result at the 5% level. Since we predict that women receive less ambitious advice, we use a one-sided test.

H3a: Treatment Effects in Advice in the Subjective Treatment

Our third set of hypotheses relates to treatment differences in the gender gap in advice. A first between-subject treatment examines what happens if referees only have access to a subjective verbal self-evaluation (and no longer see the worker's guessed score or other quantitative self-evaluations). For our power calculations, we assume that advice in this treatment is no longer connected to (guessed) performance. This implicitly assumes that subjective self-evaluations are not informative or, if they are, that male and female workers write similar subjective self-evaluations. We also assume that advisers are not directly influenced by the gender of the worker, i.e., they are not intrinsically gender-biased. In practice, this implies that we randomly generate advice in such a way that its distribution resembles the distribution of advice given in the pilot in terms of advice, while removing any connection to performance and gender. We then test whether the gender difference in advice is significantly smaller in this treatment (where it is zero in expectation) than in the baseline treatment (where it is approximately 8.6pp). Table 3 presents the results: having 400 advisers in each treatment gives us a power greater than 0.90.

Table 3: Power Calculations for the Subjective Treatment

Per Treatment:		Binary Advice	Likert Advice
#Evaluations	#Advisers	Power	Power
1000 (1000)	200 (200)	0.682	0.706
2000 (2000)	400 (400)	0.910	0.923
3000 (3000)	600 (600)	0.977	0.990
3500 (3500)	700 (700)	0.989	0.994
4000 (4000)	800 (800)	0.995	0.996

Notes: The table presents results from simulations. The first column presents the number of evaluations included in each simulation sample for the baseline and (in parentheses) the subjective treatment. The second column presents the corresponding number of advisers (one per five evaluations). The two subsequent rows summarize the results of 1000 simulations, where power is calculated as the proportion of simulated samples in which the gender gap in the subjective treatment is significantly smaller than in the baseline as per a difference-in-difference test (5% level). Since we predict a smaller gender gap in the Subjective treatment, we use a one-sided test.

Two further remarks are in order. First, since we use the baseline treatment in multiple comparisons, it may be prudent to oversample this treatment. Table 4 shows that shifting observations from the Subjective to the baseline treatment only starts reducing power when more than 30% of observations are shifted this way. Second, the true treatment effect may be larger than assumed here (e.g., if advisers are gender-biased) or smaller (e.g., if men write more optimistic subjective self-evaluations). We will discuss minimum detectable effect sizes at the end of this document.

Table 4: Power Calculations for Overrepresented Baseline in the Subjective Treatment

Per Treatment:		Binary Advice	Likert Advice
#Evaluations	#Advisers	Power	Power
2000 (2000)	400 (400)	0.902	0.917
2200 (1800)	440 (360)	0.913	0.912
2400 (1600)	480 (320)	0.904	0.917
2600 (1400)	520 (280)	0.883	0.911
2800 (1200)	560 (240)	0.860	0.842
3000 (1000)	600 (200)	0.837	0.819
3200 (800)	640 (160)	0.773	0.782
3400 (600)	680 (120)	0.710	0.686
3600 (400)	720 (80)	0.578	0.555
3800 (200)	760 (40)	0.377	0.375

Notes: See the notes to Table 3

H3b: Treatment Effects in Advice in the Nudge Treatment

Our second between-subject treatment manipulation concerns a nudging/information intervention that attempts to lower the gender gap in advice by nudging the advisers to take the existing gender gap in self-promotion into account in their advice. For our power calculations, we assume that the intervention completely eliminates the gender gap in self-confidence/self-promotion (which is equivalent to assuming that the advisers remove the gender bias in worker confidence). We implement this in the power calculation by randomly drawing observations (i.e., confidence elicitations) from both male and female workers in the pilot, and then assigning a gender randomly. This removes the gender gap in worker self-promotions by construction, while ensuring that the confidence distribution is the same as in the pilot. Table 5 presents the results, which are very similar to the Subjective treatment (Table 3).

Table 5: Power Calculations for the Nudge Treatment

Per Treatment:		Binary Advice	Likert Advice
#Evaluations	#Advisers	Power	Power
1000 (1000)	200 (200)	0.682	0.759
2000 (2000)	400 (400)	0.909	0.951
3000 (3000)	600 (600)	0.910	0.993
3500 (3500)	700 (700)	0.993	0.996
4000 (4000)	800 (800)	0.994	0.999

Notes: See the notes to Table 3

H3c: Treatment Effects in Advice in the Objective Treatment

As a final step, we include a within-subject treatment testing whether providing objective information (the worker's actual score) decreases the gender gap. For our power calculations, we simulate the data for this treatment in the same manner as in the baseline treatment above (see the section H2), except that we replace the "guessed score" with the "actual score" in the process generating simulated advice. Since we use a within-subject design, we assume that the noise term used in the simulations is identical for the baseline and Objective treatment. This essentially assumes that the adviser's reading of the worker's portfolio does not change other than through replacing the worker's subjective performance with their true performance. Table 6 shows that in this treatment the gender gap reduces by about two thirds (approximately 5.3pp on the binary scale and 0.15 points on the Likert scale), which is because the gender gap in performance (while not zero) is smaller than the gender gap in confidence/self-promotions.

Table 6: Power Calculations for the Objective Treatment

		Binary Advice		Likert Advice	
#Evaluations	#Advisers	Treat Effect	Power	Treat Effect	Power
600	120	-0.054	0.505	-0.148	0.785
800	150	-0.054	0.584	-0.153	0.882
1000	200	-0.051	0.641	-0.151	0.939
2000	400	-0.052	0.901	-0.150	0.993
3000	600	-0.053	0.971	-0.153	1

Notes: The table presents results from five sets of simulations that vary the sample size from 600 to 3000 portfolio evaluations. Binary advice is a dummy variable for whether advice A is given, and Likert advice is the five-point scale ranging from "definitely B" to "definitely A". Each row presents the averages from 1000 simulated samples. "Treat Effect" is the estimated difference-in-difference between the gender gap in the baseline and the gender gap in the objective treatment averaged across all simulation samples. Power is the fraction of simulated samples in which an OLS of the difference between the outcome in the two treatments on gender produced a significant result at the 5% level (this is equivalent to a difference-in-difference test). Since we predict a smaller gender gap in the objective treatment, we use a one-sided test.

Table 6 also suggests that a relatively small sample size may be sufficient to detect these effects. This is because we use a within-subject design and because we assume that all advisers update in line with the new information (e.g., less ambitious advice if the true objective performance was less than the subjective performance). In practice, power may be lower if advisers choose not to incorporate the objective information (e.g., because they are happy with their original advice), or if the noise term is not perfectly correlated across the baseline and objective evaluations (e.g., if advisers change their mind about the interpretation of the portfolio from the baseline to objective choice, or if they are making mistakes or are just noisy). Nevertheless, even in such cases we can expect our power to be high for the intended sample size of our experiment, as we discuss in the next section.

Final Sample Size and Cost Calculations

The previous sections suggest that a sample of around 300 workers would be sufficient to detect a gender gap in confidence with a power 0.90 (Table 1). To detect a gender gap in advice with a power of 0.90 a sample of 200 advisers would be sufficient (Table 2). And for studying treatment effects in the gender gap in advice, approximately 400 advisers per treatment would also achieve a power of 0.90. A larger sample size may still be helpful, however, in particular if the treatment effects are smaller than assumed in the power calculations. In addition, it is useful to oversample the baseline treatment, since it will be used in all of our tests (in contrast to the other treatments, which are used in a single comparison only).

Also keeping in mind budget constraints, we therefore aim to collect data from 350 workers and 1750 advisers (700 in the baseline, and 525 in the Nudge and Subjective treatments). The total costs are estimated to be around 9 AUD per adviser and 21.5 AUD per worker for a total of 23275 AUD. This also implies that each adviser has a 20% chance of being matched to a worker.

With the intended sample size, we have a power greater than 0.90 to detect a gender difference in confidence among workers based on the pilot data (H1). For the gender gap in advice (H2) the present sample size would give us a power greater than 0.90 even if the gender gap in confidence is half the size in the pilot (or equivalently: the advisers put only half as much weight on worker confidence in their final decision as they did in the pilot). For our between-subject treatments, we expect to still have a power of 0.84 on the Likert scale variable to detect a decrease in the gender gap by 60%, assuming that the gender gap in the baseline is in line with the pilot (and a power greater than 0.90 if the gender gap decreases by 70% or more).

References

Exley, C. L., & Kessler, J. B. (2022). The Gender Gap in Self-Promotion. *The Quarterly Journal of Economics*, *137*(3), 1345–1381. https://doi.org/10.1093/qje/qjac003