Pre-Analysis Plan

For "Gender Differences in Self-Promotion and Career Advice" by Rigissa Megalokonomou, Juliana Silva Goncalves and Roel van Veldhuizen. Part 2.

A. Introduction

This pre-analysis plan concerns a second data collection wave that builds upon a first wave of data collected in summer 2024. The purpose of the wave two data collection effort is two-fold. First, we want to replicate our original baseline treatment with a novel sample of advisers with teaching or management experience to test whether our results hold up in this sample. Second, we append several vignettes to the end of the survey to test for gender differences in advice received in those vignettes.

The main structure of the experiment remains the same as in our original data collection. In particular, we will 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 six workers and advises each worker whether they should go for an advanced (A) or a basic (B) version of a future task. The first five pieces of advice pertain to workers from the first wave, with the last worker being a worker from wave 2. The advisers then go through twelve vignette advice tasks with variation in (a) the confidence of the advisee (high or low), (b) the advisee's gender (male or female) and (c) the context (six different scenarios). In **session 3**, each worker receives the advice from one adviser and decides which of the two versions of the task to pursue. We will conduct only our "baseline" treatment from wave 1 in which the adviser in session 2 receives the full portfolio of worker confidence measures from session 1.

B. Main Outcomes

1. Incentivized Advice (Session 2)
    a. Binary advice (A or B)
    b. Five-point scale advice (Definitely A to Definitely B)
2. Vignette Advice (Session 2)
    a. Binary advice (A or B)

C. Secondary Outcomes

3. 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 7-great).
    g. Verbal self-evaluation

Commented [JSG1]: Should we explain that we will use the 350 workers from our previous experiment and add 105 workers to incentivise advice in session 2?

Commented [JSG2]: in our pilot, each participant saw 12 vignettes out of 24 possible vignettes (6 scenarios x gender x confidence [high vs low]). Should we have 6 vignettes only for the adviser data?

Commented [Rv3R2]: I wrote six because I did not recall the final number, but 12 is fine too! But then how do we randomize things? Do we do one of these dimensions between-subject?

Commented [JSG4R2]: We have 24 possible vignettes (6 scenarios x gender (w / m) x confidence (c / u)).
In the pilot each participant saw 12 vignettes, randomly selected out of the 24.
I imposed the following rules:
1. each participant sees 3 wc, 3 wu, 3 mc and 3 mu.
2. rounds 1–6 show first occurrence of each scenario; rounds 7–12 show second occurrence in the same scenario order (to avoid that participants see the same scenario twice in a row)
3. no scenario×gender repeats (e.g., if someone sees Jane is confident in scenario 1, they will not see Jane is not confident in scenario 1)
4. participants never see a given vignette more than once
5. Exact number of evaluations per vignette (in the pilot, 50)

Commented [JSG5]: Should this be listed as an outcome if we are analysing previously collected worker data?

Commented [Rv6R5]: Good point. I guess what I had in mind is that we would do this with the new data but perhaps this is pointless.

I am a little conflicted. On the one hand, it would be a bit strange to not analyze the 105 workers. On the other hand, it would be strange to analyse them in isolation, or to add them to the original pool. But thinking it over perhaps removing this analysis makes most sense. Or alternatively, we could say that we will include the 105 workers as a robustness check to our original main analysis of the gender gap in self-promotions?

Commented [Rv7]: I noticed that we do not control for this variable in Table 4, even though we do elicit it.

Commented [JSG8R7]: That's because advisers do not see it (they see everything else). We decided on this because we thought this would be repetitive information and so decided to leave it out from the info given to advisers

4. Other measures of incentivized advice (Session 2)
    a. Performance target for workers
    b. Confidence in binary advice

D. Main Explanatory Variables

5. Worker Gender: dummy variable comparing males to females.
6. Confident Vignette: dummy variable comparing vignettes with a confident worker (1) to vignettes with a non-confident worker (0).
7. Vignette scenarios: six dummy variables equal to 1 for choices made under one of the six vignette scenarios.

E. Other Variables

8. Advisors: gender, other demographics, confidence/risk aversion
9. Advisor beliefs about gender differences:
    h. Are men or women better at the task?
    i. Are men or women more confident in their self-assessment?

Demographics include personal income, household income satisfaction, age, occupation, job title, education, experience in STEM fields, experience in teaching and / or management, experience in giving advice to others in the context of their teaching and / or management experience.

F. Main analysis

We will include in our main analysis only the first five workers evaluated by the advisers. These workers are the same exact workers our advisers encountered in wave 1, allowing us to attribute any differences we observe between waves to advisers rather than workers. We will conduct robustness checks including the sixth worker as well (G1 below). We use one-sided tests for gender gaps in advice received (F1 and F3), since we expect women to receive less ambitious advice based on both previous research and our results from wave 1. We will use robust standard errors clustered at the adviser level for any regression analysis unless otherwise specified.

Our experiment aims to answer the following research questions:

1. Gender Gap in Advice (Incentivized Task): Do female workers receive less ambitious advice than male workers?
    a. Main test: two linear regressions of the (a) binary advice measure and (b) Likert scale advice measure respectively on a dummy for the worker's gender. We use a one-sided test on the gender dummy.
    b. With controls: two linear regressions of the (a) binary advice measure and (b) Likert scale advice measure respectively on a dummy for the worker's gender. We control linearly for each of the worker confidence variables C3a-e. We use a one-sided test on the gender dummy.
2. Experience Effect on Advice: Is the gender gap similar in the wave 1 sample and the wave 2 sample with advisors with management/teaching experience?
    a. Main test: a linear regression of advice on gender, experience, and a gender*experienced interaction. The key test is a two-sided test on the significance of the gender*experienced interaction. Here, "experienced" is a dummy equal to one for participants from the second wave (those with management/teaching experience).
    b. With controls: same as the main test but now controlling linearly for each of the worker confidence variables C3a-e.
    c. With interacted controls: same as (b) except we now interact all confidence variables C3a-e with the experienced dummy.

Commented [Rv9]: I'll need to revisit the numbering elsewhere after removing part 3 above. (References might now be incorrect.

Commented [Rv10]: 1 and 2 are pretty much the same as before, except that 2 is about experience rather than treatment effects. The main difference is that the latter is a two-sided test because I, at least, did not have a clear hypothesis for how the experience effect should go...

Commented [JSG11R10]: I agree

Commented [Rv12]: To allow for experienced advisers to be more/less influenced by worker confidence. Could also be a robustness check.

In addition to investigating the incentivized measure, we will also examine the following hypothesis related to our vignettes:

3. Gender Gap in Advice (Vignette Task): Do women receive less ambitious advice than men?
    a. Overall: linear regression of the vignette advice measure on a dummy for the worker's gender. We use a one-sided test on the gender dummy. For this analysis, we will pool data from all vignettes and include scenario-fixed effects (one for each of the six possible scenarios D7) and control for worker confidence D6 (1-confident, 0-not confident).
    b. By vignette: we repeat the previous test separately for each specific scenario. This gives us the gender gap in advice separately for each scenario.
    c. By confidence: we repeat test (a) but include a confidence*gender interaction effect. We use a one-sided test for both the gender dummy (the gender gap among non-confident workers) and the sum of the gender dummy and interaction term (the gap among confident workers). For the interaction term, we will use a two-sided test.
4. Comparison between Vignettes and Incentivized Task: is the gender gap similar in the two tasks?
    a. Main test: Regression comparison: linear regression of binary advice on a dummy for the worker's gender, a dummy for vignettes (1-vignette, 0-incentivized) and their interaction. We use the data from all vignettes and include scenario-fixed effects for the vignettes. The key test is a two-sided test on the interaction term.
    b. With controls: same as the main test but now controlling linearly for the confidence variables C3a-e (incentivized data) and for the confident vignette dummy D6 (vignette data).

G. Robustness Checks

We plan to include the following robustness checks.

1. Main results with all workers: do our main results F1, F2 and F4 hold up when also including the sixth worker (who only received advice in wave 2)?
    a. Replication of analysis F1, F2 and F4 while now also including the data for the sixth worker that receives advice.
2. Gender Gap in Self-Promotion: Are female workers are less confident than male workers?
    a. New sample: one-sided t-test for each of the quantitative worker confidence variables listed in part C3 above (a-g) for the 105 workers in wave 2.
    b. New sample with controls: Regression of confidence on the worker gender dummy (1-female, 0-male) while controlling for worker score fixed effects from session 1. The relevant test is a one-sided test on the regression coefficient on the female dummy. Separate test for each variable listed in part C3 (a-g) above. 105 workers from wave 2 only.
    c. Full sample: same as (a) but now including all workers from wave 1 and wave 2.
    d. Full sample with controls: same as (b) but now including all workers from wave 1 and wave 2.
3. Are the gender gap in vignette advice F3 and the difference between vignettes and the incentivized task F4 robust to only looking at the first six vignettes, i.e., to only encountering each scenario once?
    a. Replication of main analysis F3 and F4 while including only the first six vignettes evaluated.
4. Main results with first worker: do our main results F1, F2 and F4 hold up when only analyzing the first worker?
    a. Replication of analysis F1, F2 and F4 while only including the first piece of advice.
5. Do gender differences and experience effects in advice (F1 and F2) also appear in our secondary measures (C4)?
    a. Replication of main analysis F1 and F2 for the performance target C4a.

Commented [Rv13]: In our pilot sometimes the gap was reversed. Nevertheless I'd still pre-register the one-sided test. We could always mention the two-sided result if it goes in the other direction in the paper. We could also pre-registered b-d as additional analysis rather than main analysis.

Commented [JSG14R13]: a. is equivalent to column 1 Table 4 in the paper (where we test for a gender gap w/o additional controls). Why don't we have a regression that controls for confidence (similar to column 2 Table 4)? c. is interesting, but we do not test this with the experimental data. But that's probably not an issue. In that case, should we have an interaction term instead or running 2 separate regressions? (confidence x gender). Same for d. - should we interaction confidence x gender for each vignette scenario?

Commented [Rv15R13]: My replies:
a.I think it is a good idea to control for confidence. The difference is that previously confidence was elicited, now it is a treatment variable. So what I did for now is to add a dummy for worker confidence. A separate specification could work too but might not be very interesting.
c+d. Yes an interaction term makes sense as a formal test.

We could also do fewer tests here, or leave some of the tests for additional analysis.

Commented [Rv16]: We discussed having scenario*gender interactions here instead. I actually was about to implement this but then I realized that we don't have a logical "baseline" scenario here. We could of course re-run this regression with any of the six scenarios as the baseline, but that would be too much. So in the end, running each regression separately for each scenario still made the most sense to me.

Commented [JSG17R16]: I agree, that's a good point

Commented [JSG18]: Can we include vignette FE, since we would not have this in the experimental data?

Commented [Rv19R18]: I am not sure, let's discuss. Also fine with removing them.

Commented [JSG20]: that seems more straightforward than I was thinking, which is good :)
However, in the experimental data, we expect a gender gap because women are less confident. In the vignettes, there is a gender balance in confidence. That means we expect ...

Commented [Rv21R20]: I am also fine predicting a smaller gender gap in the vignettes because of similar confidence. However, it could also be larger there because it is a more natural setting, may activate stereotypes more, etc. ...

Commented [JSG22]: since we are planning to mainly use the previous worker data (N=350) and add 105 new workers for incentivisation of adviser decisions, I am not sure 1-3 are relevant. I thought our analysis would also focus on the workers from the previous data collection, to maintain the ...

Commented [Rv23R22]: I agree. We could either delete this or replace it with a full analysis including all workers as a robustness check to what we already have.

   b. Replication of main analysis F1 and F2 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.
6. Do experienced advisers give better advice?
   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 advice on the "experienced" dummy (see F2 above). Our test is a one-sided test on this dummy.
   b. We regress expected earnings based on following the advice and part 1 score on the experienced dummy. Our test is a one-sided test on this dummy.
   c. We create a new variable to compare the gap in expected earnings based on advice received and what would have maximized expected earnings based on part 1 score. We regress this variable on the experienced dummy. Our test is a one-sided test on this dummy.
   d. We will also repeat regressions a-c while controlling for gender and interacting gender with the experienced dummy. This allows us to test whether the gender gap in advice quality is different for experienced advisers. Here the key test is a two-sided test on the gender*experienced dummy.

H. Sample, Recruitment and Randomization

We aim to recruit 525 advisers and 105 workers from the sample of UK participants on Prolific. We will require adviser participants to have experience in either management or teaching roles. Should we fail to reach our target sample of advisers within one week, we will supplement our sample with Americans using the same criteria. If more than 20% of our adviser sample ends up being US-based, we will conduct a robustness check for our main analysis (section F) using only our UK advisers.

Advisers first receive portfolios from five workers from our first data collection effort (wave 1). Each of the 350 worker portfolios from wave 1 gets randomly assigned to either 7 or 8 advisers. Each adviser then receives the portfolio from one of the workers from wave 2, so that each worker's portfolio is randomly assigned to five unique advisers.

For the vignette task, the first six vignettes the advisers see correspond to the six scenarios, presented in random order. For each of these six vignettes, we randomly vary the worker's confidence level (high or low) and gender (male or female). They will then see the same six scenarios a second time, changed so that either confidence or gender is different from the first set of six scenarios.

Power Calculations

Summary

We conducted power calculations to check our power for the four components of our main analysis, that is, to detect (a) a gender gap in incentivized advice, (b) the experience effect on incentivized advice, (c) a gender gap in vignette advice and (d) comparisons between vignette and incentivized advice. Based on these power calculations, we aim to collect data from 525 advisers and (for procedural reasons) 105 workers. With the intended sample size, we will have (a) a power greater than 0.80 to detect a gender gap in incentivized or vignette advice of 40% the size of the gap in our first wave and (b) a power greater than 0.80 to detect a 60% difference in advice due to experience or due to the task used (vignettes vs incentivized tasks). In the remainder of this document, we will present the power calculations in greater detail.

Commented [JSG24]: We only need 105 workers. If each adviser has a 20% their advice is sent to a "new" worker, we can achieve this with 105 workers

Commented [Rv25R24]: This was a typo. The workers are anyway not necessary for any power calculations.

H1: Gender Difference in Incentivized Advice

Our first hypothesis relates to the gender gap in advice. In our first wave we collected data from 350 workers and 700 baseline treatment advisers who each evaluated 5 worker portfolios for a total of 3500 adviser-level observations. We compute power for the binary advice measure in wave 2 by using simulations to randomly draw A-advice using the first wave averages for men (43.6%) and women (31.5%). We do this for our intended sample size (525 advisers and 2625 evaluations) and calculate power for effect sizes equal to or smaller than in our first wave. For the Likert scale we use a similar approach where we draw advice equal to the average response for men (2.75 points) and women (2.37 points) respectively and then add noise to achieve a similar R^2 as in the first wave data set. Table 1 presents the results. The main takeaway is that we still have good power (>.80) even for a gender gap only 40% of the size of our first wave (-0.048 in the binary measure).

Commented [JSG26]: if we only use the first 5 workers to keep the exact same sample of workers, we would have 2625 obs. Do we want to keep the same sample of workers (to focus on advisers only and comparison of adviser type across treatments)?

Commented [Rv27R26]: Yes, a good idea, forgot about this feature. I've updated the power calculations accordingly.

Table 1: Power Calculations for a Gender Difference in Advice

			Binary Advice		Likert Advice	
	#Evaluations	#Advisers	Gender Gap	Power	Gender Gap	Power
100%	3150	525	-0.121	1	-0.38	1
80%	3150	525	-0.097	1	-0.30	1
60%	3150	525	-0.072	0.98	-0.22	0.99
40%	3150	525	-0.048	0.81	-0.15	0.85

Notes: The table presents results from four sets of simulations that vary the effect size from 100% to 40% of the first wave effect in steps of 20pp. 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.

H2: Effect of Experience

Our second hypothesis relates to the effect of management/teaching experience. To calculate power, we use the same simulations as for hypothesis 1 and compare the simulated gender gaps to the observed gender gap in our first wave data. Here the key test is a two-sided t-test on the interaction term of a regression of advice on gender, a dummy for experienced (i.e., wave 2) advisers and their interaction. The results (presented in Table 2) suggest that with our intended sample size we have a power of at least 0.83 to detect 60% increases or 60% decreases in the gender gap, with a larger power for larger changes.

Table 2: Power Calculations for Experience Effect

	Binary Advice			Likert Advice		
	Baseline	Experience	Power	Baseline	Experience	Power
+100%	-0.121	-0.242	1	-0.38	-0.72	1
+60%	-0.121	-0.193	0.92	-0.38	-0.58	0.83
+50%	-0.121	-0.182	0.76	-0.38	-0.55	0.65
-50%	-0.121	-0.059	0.72	-0.38	-0.19	0.81
-60%	-0.121	-0.048	0.91	-0.38	-0.15	0.92
-100%	-0.121	0	1	-0.38	0	1

Notes: The table presents results from six sets of simulations that vary the gender gap among experienced advisers. Power is the fraction of simulated samples in which an OLS regression of the outcome on gender, a dummy for experienced adviser and their interaction produced a significant interaction term at the 5% level. See the notes to Table 1 for further details. We use a two-sided test since we have no ex ante prediction of a different gender gap among experienced advisers.

H3: Gender Gap in Vignette Advice

Our third hypothesis posits that women receive less ambitious advice in the vignette tasks. In our pilot, 55% of men received ambitious advice across all vignettes. With our intended sample size of 525 evaluators and 6300 evaluations, we can calculate minimum effect size using the power twoproportions command in Stata. If we assume that each piece of advice is an independent observation, we obtain a minimum detectable effect size of 3.6pp to detect a gender gap across all vignettes with a power of 0.90 (power twoproportions 0.55, n(6300) power(0.9) alpha(0.05) onesided). The minimum detectable effect sizes are 5.2pp when split by confidence level (n=3150 evaluations), 8.9pp for each individual vignette scenario (N=1050) and 12.4pp for each vignette/confidence combination (N=525). If we instead assume that pieces of advice are not independent but correlated for each adviser, our minimum detectable effect size would lie somewhere between 12.4pp and the respective lower values mentioned above.

Commented [JSG28]: In the pilot, each participant saw 12 vignettes. Do we want to reduce it to 6?

Commented [Rv29R28]: I didnt remember the number. 12 would give us even more power.

H4: Comparison between Vignettes and Incentivized Advice

Our fourth hypothesis compares the gender gap between vignettes and incentivized advice. For this purpose, we can use simulations similar to H2 above. For the incentivized data, we simulate new data assuming the same rates of ambitious advice for men and women as in wave 1 in the baseline (2625 evaluations). For the vignette task, we use the same process, where we vary the size of the gender gap to be smaller or larger than in the incentivized task (6300 evaluations). Power is then calculated as the fraction of simulated

samples in which we observe a significant interaction term in a regression of advice-A on gender, a vignette dummy and their interaction. Table 3 shows that we have good power to detect a 50-60% increase or decrease in the gender gap between the vignettes and the incentivized task.

Table 3 Power Calculations for Vignette vs Incentivized Task

	Incentivized	Binary Advice Difference	Power
-100%	-0.121	-0.121	1
-60%	-0.121	-0.073	0.91
-50%	-0.121	-0.060	0.78
+50%	-0.121	+0.059	0.77
+60%	-0.121	+0.071	0.91
+100%	-0.121	+0.121	1

Notes: The table presents results from six sets of simulations that vary the gender gap in the vignettes while keeping the gap in the incentivized choices constant. Each simulated sample contains 2525 incentivized evaluations and 6300 vignette evaluations, both split equally across genders. Incentivized is the gender gap among incentivized evaluations. Difference is the difference between the incentivized gap and the vignette gap in the simulations. Power is the fraction of simulated samples in which a significantly different gender gap is observed between vignette and incentivized responses (two-sided test).