Australian Government
Department of the Prime Minister and Cabinet

BETA

Evaluation options for a job ads trial

This document sets out three different evaluation options that test the effect of changing the language and requirements in job ads on the gender balance of applicants. It complements a project we completed in collaboration with the Department of Industry, Science and Resources (DISR), *Attracting a Diverse Cyber Security Workforce: Lessons from an Analysis of Australian Job Ads*, in which we found that reducing gendered language[1] in job ads could encourage more women to apply. The purpose of this document is to outline the benefits and risks of running a trial on job ads, as well as the feasibility of any trial. It is based on BETA's collective experience and knowledge and represents initial thoughts only: Any actual trial would require further detailed planning.

The three trial designs we describe are:

1. Posting fake job ads with gendered vs. neutral language on an online jobs board.
2. Training HR Professionals in how to reduce gendered language and providing a tool to measure it in the job ads they write.
3. Changing only the job titles of real ads.

For each of these trials, we would use different methods to randomly vary the text of the job ads, to be either more or less gendered. We would then measure the proportion of men and women who apply to each job ad. This makes these trials a little different from normal trials: The recipient of the “treatment” is the job ads, not the people applying. To make this more concrete, in our trial the job ad is receiving the “dose” (how gendered the language is), and the outcome we are measuring is the effect on the job ads’ outcomes (i.e. the gender balance of applicants). Therefore, these trials would be making inferences about the population of *job ads* in Australia, not the population of *people* in Australia.

Benefits of running trials

There are a number of benefits to running a trial in this space that are true for all of the trials. Principally, while there is some primary research in the literature, most studies to date have involved hypothetical outcome measures (e.g. “how likely would you be to apply to this job, on a scale from 1 to 7”). This trial would be one of a small list of real-world trials, and the only in Australia. This trial would provide solid evidence about the effects of specific language in

[1] Other aspects of job ads which may influence the number and proportion of women who apply also include whether the job ad mentions the availability of flexible work, and whether it includes long ‘laundry lists’ of requirements. These aspects could also be evaluated; however for the sake of simplicity in this document we focus on gendered language (i.e. language associated with masculine or feminine stereotypes).

job ads that could be used in Australia, as well as internationally. Any trial also aligns with the government's aim to reduce the level of gender inequality in Australia. A positive result would suggest that employers in male-dominated industries should work harder to improve the language in job ads in order to attract more female applicants. A null result would indicate that job ads do not have an effect on people's behaviours, and that efforts to improve gender balance in various industries should be directed elsewhere.

Costs of trials

The cost of running a trial would mostly be driven by the opportunity cost of staff time. Any trial would be time consuming and costly to run, and resources may have more use elsewhere. Based on findings from previous research we are also unlikely to see a very large effect when changing the language in job ads, and so the overall effect on gender equality may be small.

Producing the evidence also has risks. Currently, potentially sexist workplaces may be writing sexist ads – i.e. ads that use highly stereotypically masculine language (this is based on conjecture). If this is the case, women would currently be less likely to apply to these workplaces – and this would be a positive outcome. However, better evidence and advice on how to write job ads that are attractive to women might result in job ads for some workplaces becoming 'wolves in sheep's clothing' if sexist workplaces become better at advertising to women as a result. This in turn may result in female applicants being harmed, if they unknowingly apply for sexist workplaces because the job ads no longer function as an accurate signal of the values of the company. While this is a potential risk of building the evidence on gendered language in job ads, more broadly there may be broader cultural shifts that need to occur to address this.

Difficulties of running trials

For each of the three trials, reaching an appropriate sample size will be a challenge. If this trial replicates the usual effect of behavioural economics trials, we would likely only see a small treatment effect (3-6%). This would mean that we would need to post roughly 250-1000 job ads online. Given the effort required to create job ads, either the trial would need to be in the field for a long time, or it would be a costly exercise.

For any trial design, we would need to partner with a number of large private or public sector organisations. Partnering with any organisations can be difficult and take a long time. Changes in a partner organisations' hiring policies during the trial can also threaten the trial's validity.

Lastly, the sample of Australian job ads used in the trial would face selection bias, as they would be chosen (or written) by us or the partner organisation, rather than being selected randomly. This means that the generalisability of this trial would be reduced, and we would likely face voltage effects (List 2023): i.e. when scaling up, the effects are unlikely to be as large. For very small effect sizes, there may be no practical significance of the effect, even if the result of the trial was statistically significant.

Fake job ads

Overview

The aim of this trial would be to test the effect of gendered language in jobs ads on the gender balance of applicants. Following the methodology of Burn et al. (2022), we would post fake job ads on an online jobs board and then collect the demographics of applicants.

We would need to partner with a large online job board (e.g. SEEK or LinkedIn) to post ads to the website**,** as posting fake ads is against the terms of service for most organisations. We may also need to partner with real organisations or recruiters so that the job ads could include the name of real companies**.** If we use fake organisations, the job ads will look like scams, and this would potentially invalidate any inferences we make. However, posting fake job ads involves risks for all organisations involved.

Table 1. Pros and cons of a trial using fake job ads

Pros	Cons
• Low cost intervention option. • Easy to implement. • We would only need to get one organisation on board (the job board website). • The level of control we have over the job ads would mean that we would need a much smaller sample size. • We would be likely to see a bigger effect of varying the treatments compared with other trial designs.	• Ethics of companies posting fake job ads is questionable. The problem is that people may put undue effort into a job application where they have no chance of receiving a job that doesn't exist. • Risk of the government of running a trial that is inherently deceptive. • We are measuring the effect of gendered language, not how to fix it. Any recommendations would be based on our thoughts, not tested interventions. • We might not be able to find a partner in this space. Organisations like SEEK may be very resistant to posting fake job ads.

Trial aim

To test the effect of gendered language on the demographics of applicants.

Trial design

We would run this experiment as a multi-arm randomised trial. We would randomise which job ads receive the treatment, and for each group of job ads (treatment vs control) we would measure the demographics of applicants.

We would run the experiment by first creating base job ads based on real job ads (our 'control' job ads). The base job ads would be designed so that they have would be neither more nor less stereotypically gendered than the average job ad. Our intervention would be to vary the language in the job ad text. We would vary the text using phrases from real job ads so that the language is not contrived.

Interventions

Control job ads

This group would be unaltered base job ads.

Treatment job ads

The intervention would be changing the language in the job ad text. This could include removing, editing, or adding additional words or phrases to the job ads. The aim would be to make the language more or less stereotypical.

We would vary the language in a number of ways:

- **Stereotypical body text:** We would vary the body text of the job ad in three ways:
    - Increasing the presence of stereotypically masculine language. In practice we would hope to raise the similarity score by 1 standard deviation from the mean ad for a number of stereotypes.
    - Increasing the presence of stereotypically feminine language. In practice we would hope to raise the similarity score by 1 standard deviation from the mean ad for a number of stereotypes.
    - While we could look at increasing the similarity score for specific stereotypes, this may increase the sample size to much (see the power analysis).
- **Stereotypical job title:** Changing the job title. Specifically for cyber we could look at the effect of a cyber 'analyst' (a stereotypically masculine term) vs other terms.
- **Flexible work:** Vary the inclusion of flexible work in the job ad, for example including specific references to part-time, job share options.

We would test each intervention against the control, as well as the combined effects for some of the treatments as an optional inclusion. For the combination treatments we could look at the:

- ‘best’ combination of attributes: no masculine language, increased feminine language, not stereotypical, no stereotypical job title, and including flexible work.
- ‘worst’ combination of attributes: masculine language, no feminine language, stereotypical, stereotypical job title, and no flexible work.

This design allows us to include only interaction that we are interested in, as opposed to all of the different interactions of the different interventions. The inclusion of these combination arms would be dependent upon sample size as this type of testing will greatly increase the required number of job ads.

A mock-up of what this design could look like is shown in Figure 1.

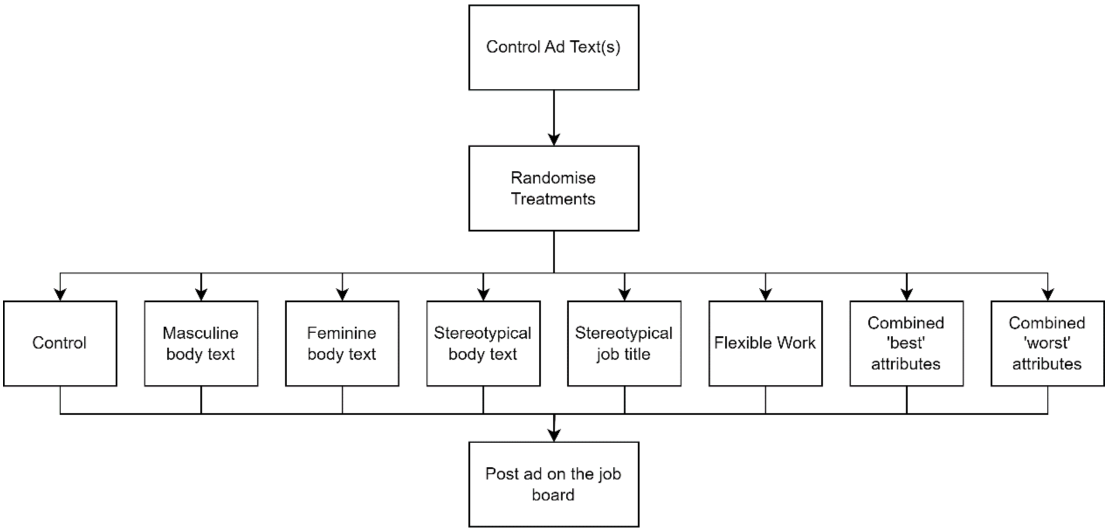

Figure 1. Diagram showing the randomisation scheme for the trial. It shows how we would randomise the treatments and covariates among the fake job ads

Outcome measures

Primary outcome

We would measure the proportion of people who identified as ‘male’ that applied to each job ad. We use ‘not men’ as the outcome of interest as we want to include people of other genders besides female.

Secondary outcome

We could also collect a number of additional demographics as secondary outcome measures including the age of participants and level of education. There are different options for how we collect this information. We could infer it based on peoples CV’s or we could ask it directly. Asking directly may, however, change the people who apply, if some people do not want to disclose this information. We would also look at the overall number of applicants for each ad.

Data collection

We would collect the demographics of applicants at the point that applicants submit their CV.

Population and sample selection

The population would be ‘job ads posted in Australia’. Our sample would be based on the ‘average’ job ad, as measured by the average characteristics we selected. This means that we may have limited external validity, as these job ads would be written by us.

An alternative approach to the trial design would to be randomly sample a dataset of job ads (such as Lightcast), and then copy the ad text, then vary the language. This would make inferences about the population of job ads more valid, but may be a breach of copyright.

Hypothesis

Individual hypothesis

- The proportion of *male applicants* would be **lower** for job ads which had *less* stereotypically masculine language compared with control.
- The proportion of *male applicants* would be the **same** for job ads which had *more* stereotypically feminine language compared with control (a null).
- The proportion of *male applicants* would be **lower** for job ads which had *less* stereotypical language compared with control.
- The proportion of *male applicants* would be **lower** for job ads which offered flexible work compared with control.
- The proportion of *male applicants* would be **lower** for job ads that that included titles that are less gendered compared with control.

Combined hypothesis

- The proportion of *male applicants* would be **lower** for job ads that combined the most feminine attributes (the 'best' option) compared with all other treatments. This would be a conjunction test.
- The proportion of *male applicants* would be **higher** for job ads that combined the least feminine attributes (the 'worst' option) compared with all other treatments.

In these cases we will only reject the null for the joint hypothesis if we reject the null for all constituent hypotheses.

Randomisation

Randomisation would occur at the job ad level. Job ads would be allocated to one of the 7 treatments or the control with equal probability.

Sample size and power

Burn et al. showed that including 3 ageist machine learning phrases reduced the proportion of older applicants by 12 percentage points, changing the proportion of older applicants from roughly 19% to 31%. The proposed trial has a similar design and may achieve a similar effect size. However, typically such trials have much smaller effects.

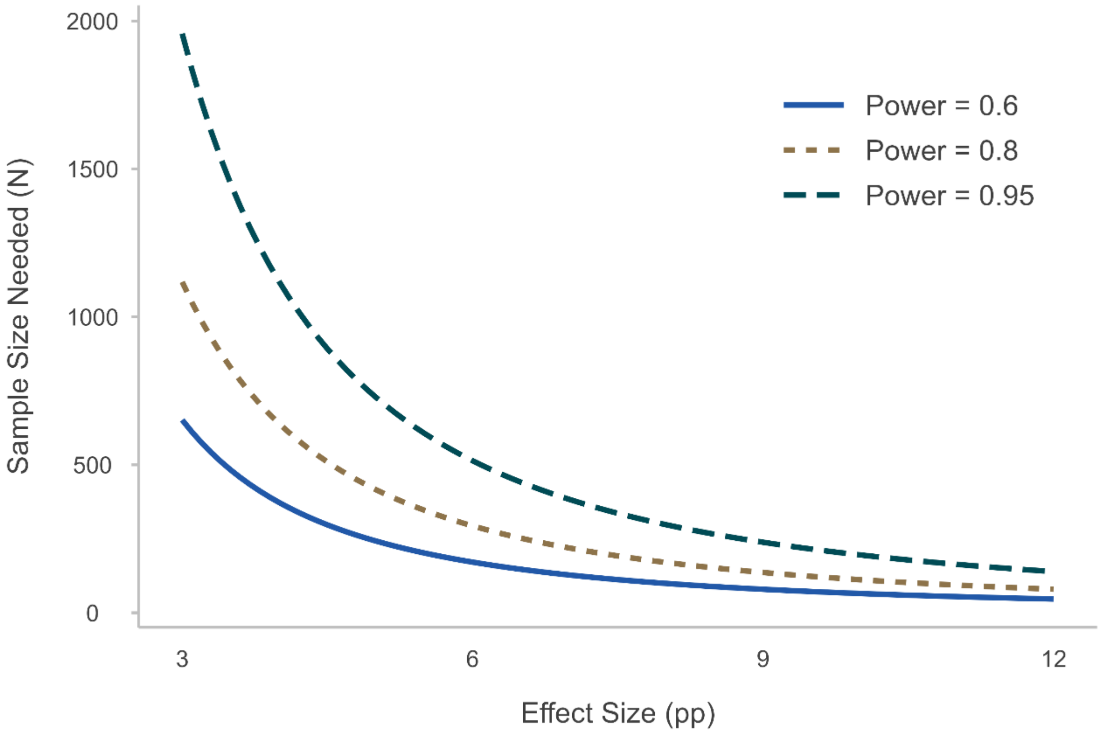

Figure 2. Power analysis for the fake job ads trial, varying the effect and sample size.

Figure 2 shows that there is a large range of sample sizes needed given the effect size we could feasibly get. If we get roughly half the effect of the Burn et al. (2022) results we could get reasonable power from roughly 300 ads per arm of the trial. This would mean a minimum trial size of 2400 job ads (7 interventions and one control). The conjunction test would mean that we would need a larger sample size. In the full trial we would estimate the power for this design through simulation to get a more accurate measure.

Training people to reduce gendered language

Overview

In this trial we would partner with a number of HR teams to test the effectiveness of both training and a tool to make the language in job ads less gendered, and thus increase the proportion of women who apply. To do so, we would need to partner with the recruitment teams from several organisation in order to administer the training and give access to the tool. For each organisation we would randomly allocate HR professionals to receive the training and be able to use the tool (or not). We would then measure the gender balance of the applicants for each job ad, and see whether more women applied to job ads written by HR professionals who received the intervention (than to ones written by HR professionals in the control group).

Table 2. Pros and cons of a trial that involves training HR professionals

Pros	Cons
• We would be testing the effectiveness of an intervention to reduce gender bias. This is more useful for organisations as they can more easily apply the findings from this trial. • We could measure whether the training impacted the overall hiring outcome.	• It may be difficult to find enough organisations to run a powered trial. • The impact on the gender ratio of applicants may be small. • It would be difficult to stop cross-contamination of the treatment. That is, HR professionals who received the training telling their colleagues to change the style of their ads. • It would be an expensive trial in terms of time/cost of materials to get the training set up and to implement the intervention. • In BETA's experience HR data is difficult to work with.

Trial Aim

To test whether a combination of training and a tool to improve the way people write job ads increases the proportion of women that apply.

Trial Design

The design of the trial would be a stratified, clustered randomised controlled trial, as illustrated in Figure 3. We would randomly allocate each HR professional to the treatment, stratifying by each organisation. We would then randomly allocate the job ads that each HR professional received (to reduce bias), and for each job ad, measure the demographics of applicants. To see the effect of the intervention we would then compare the difference in the demographics of applicants applying to job ads written by the group that received the intervention, compared to those written by the group that didn't.

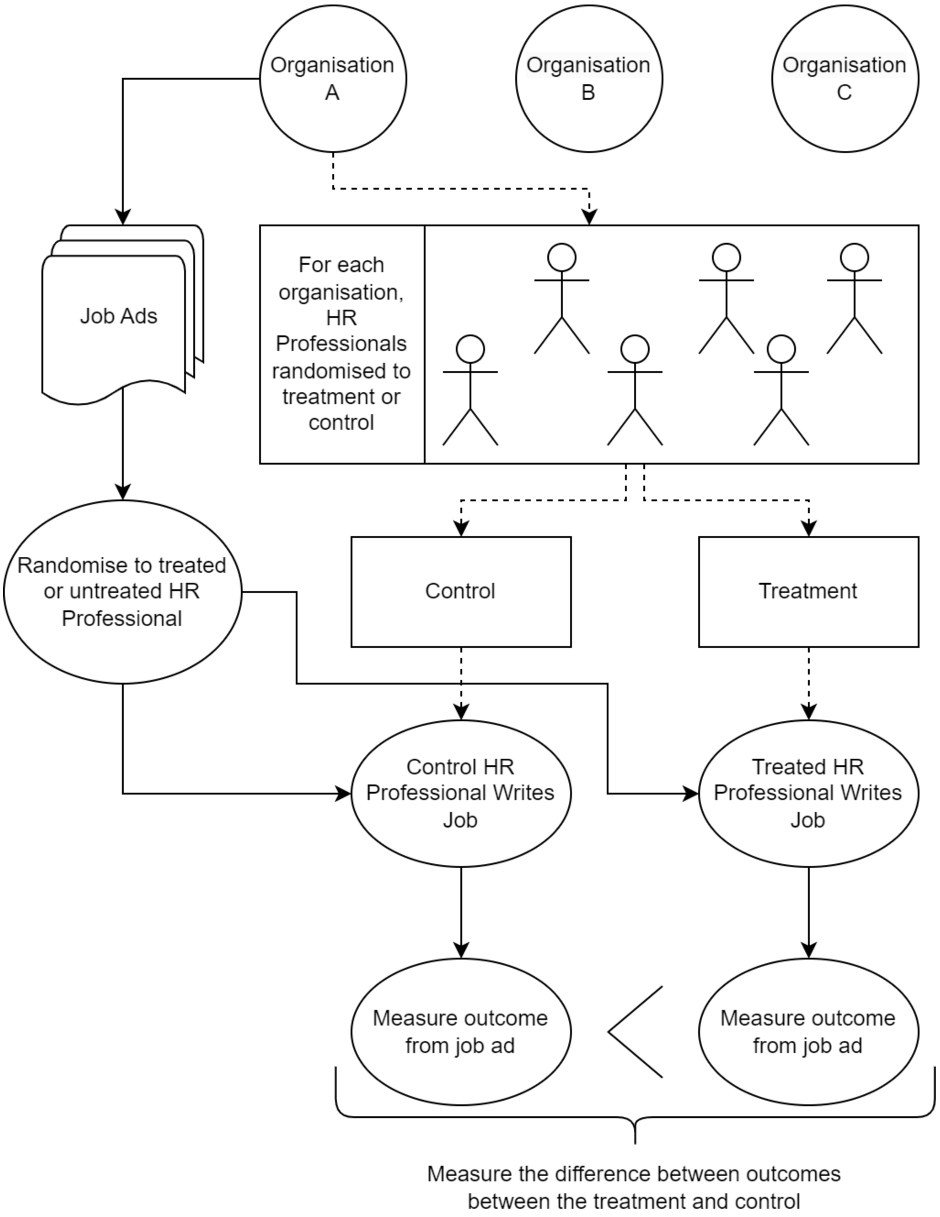
Organisation A
Organisation B
Organisation C
Job Ads
For each organisation, HR Professionals randomised to treatment or control
Randomise to treated or untreated HR Professional
Control
Treatment
Control HR Professional Writes Job
Treated HR Professional Writes Job
Measure outcome from job ad
Measure outcome from job ad
Measure the difference between outcomes between the treatment and control

Figure 3. Diagram of the intervention design for the HR Trial. It shows how we would randomise the treatment for HR professionals, as well as randomise which job ad each HR Professional writes.

Intervention

We would only test the effectiveness of the training and the tool in combination because this would probably give us a stronger effect on the language used in the job ad text than either component individually.

The control group would write job ads business as usual.

Data collection strategies

We would rely on each partner organisation to collect demographics of applicants, at the point that applicants submit their CV. This may require a change in the HR systems of the organisations.

Primary outcome

The primary outcome is the proportion of male applicants who applied for each ad. To measure this, we would collect demographic information of applicants in the first round of applying (e.g. at the CV submission/cover letter).

Secondary outcomes

There are a number of secondary outcomes we could consider:

- Other demographics collected at the initial stage of the trial, e.g. age, and CALD, indigenous and disability status. Given the language is more inclusive we would expect that there may be crossover effects of the intervention.
- The demographics of the people successfully employed.
- The number of applicants in each group.

We would also collect the characteristics of each job ad (level of seniority, pay) to include as covariates in the analysis.

Population and sample selection

The population would be job ads written in Australia. However, our sample would be biased to the organisations that agreed to partner with us. This may not be generalizable to all job ads in Australia, because of selection bias.

Hypothesis

The proportion of *male applicants* would be **lower** for job ads written by HR professionals in the treatment group compared with the control.

Randomisation

Randomisation would be stratified by the different partner organisations. Individual HR professionals would be randomised within each organisation to receive the treatment or not. This is a cluster based design as each HR Professional is ‘treating’ each ad. To reduce the intra-class correlation coefficient, we would randomise which job ad each professional received.

Sample and power analysis

There are a number of factors that can affect the sample size needed in order to have a well powered trial. We considered:

– the number of organisations we could sign up,
– the size of their HR teams,
– the intra-class correlation coefficient (ICC) of each HR professional,
– and the number of ads that each individual would need to see to get a powered trial.

We present the different powered trials below (Figures 4 to 6). The power analysis shows a large variation in the power for a trial based on the effect size and number of organisations.

Given this, an under-powered trial is a specific risk to this trial design. To detect 4 percentage point effect size, in the best situation we would need to sign up roughly 6 organisations with approximately 15 HR professionals posting 15 job ads. The trial would need to be in the field for roughly 6 months to achieve sufficient job ads. However, these are indicative only and when undertaking a trial a full sample size analysis with specific organisational data would be needed.

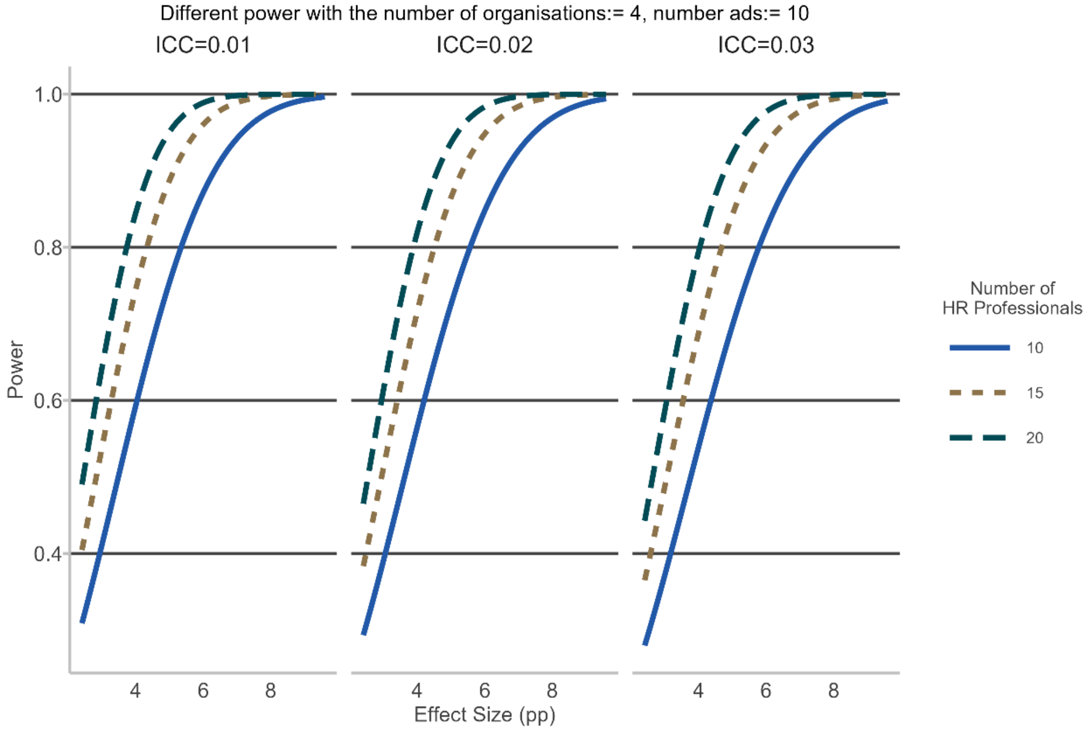

Figure 4. Power analysis for a fixed number of organisations, and job ads, while we vary the number of HR professionals and the ICC for each ad. Varying the number of HR professionals and job ads has a similar effect (see Figure 5).

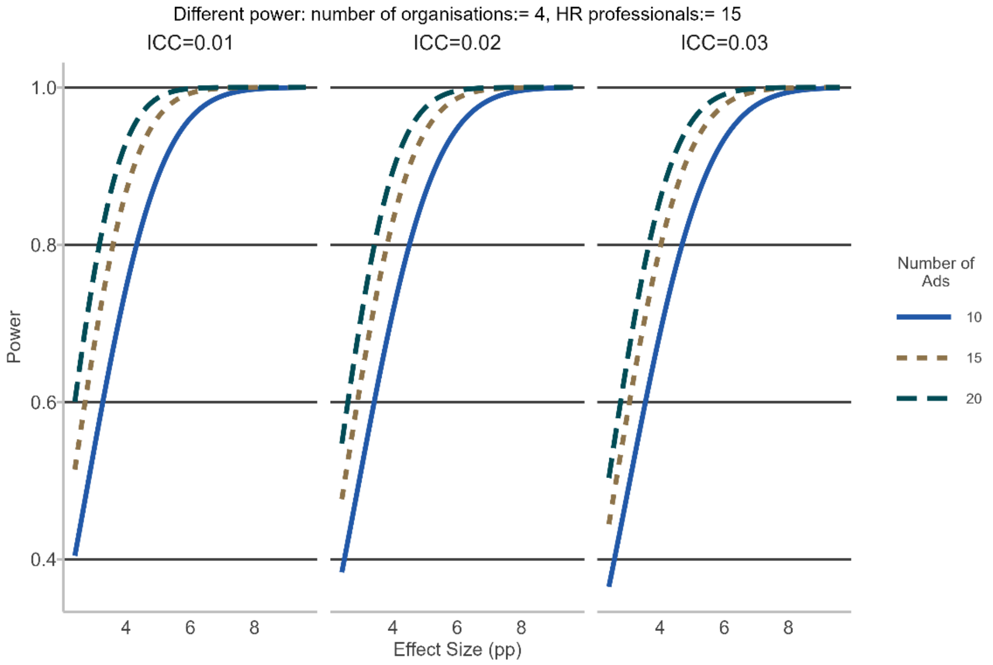

Figure 5. Power analysis for a fixed number of HR Professionals and organisations while varying the number of job ads. This has a smaller effect on the power.

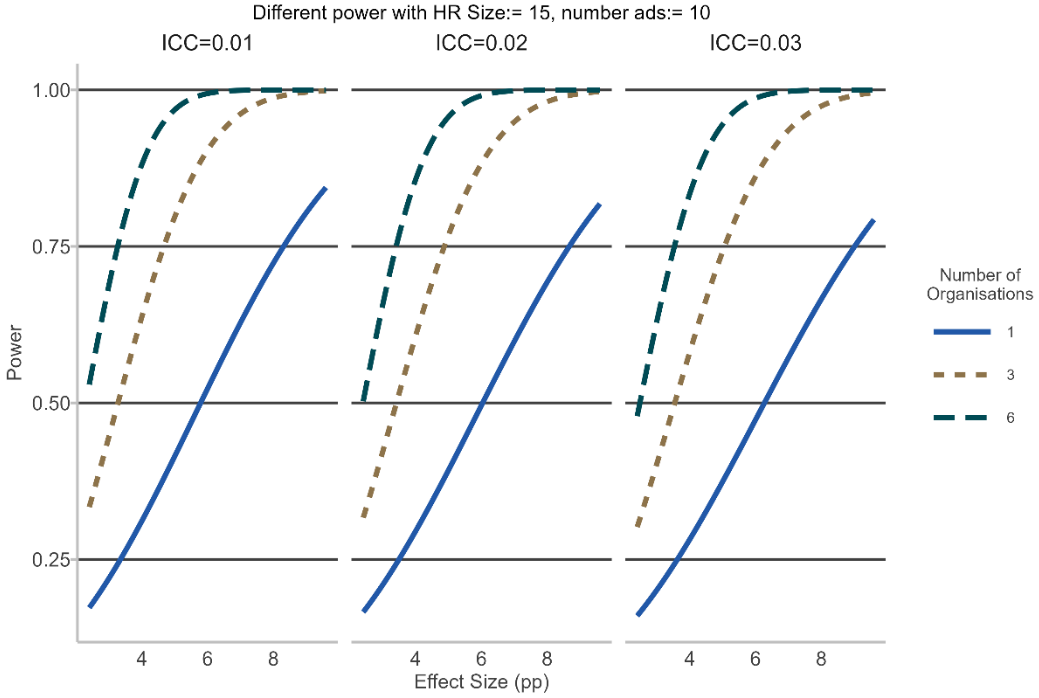

Figure 6. Power analysis for a fixed number of job ads and HR professionals, while varying the number of organisation. This has the biggest effect on the power (cf Figures 4 and 5).

Trial threats

There are a number of threats to the validity of this type of trial:

- Cross-contamination of the treatment variable is a real risk with this trial. If HR professionals share the information about how to write ads more effectively it is a risk to the validity of the trial. We may be able to see this if we measured how gendered the language is in each job ad, and could show that they changed over time even in the control group
- If we fail to randomise which ads go to which person, the trial design may be invalid.
- The possibility that we get drop out is a threat to the trial. If HR professionals leave the trial due to normal attrition (from their organisation) then it could be difficult to reach statistical power. We could address this by planning on randomly allocating people to the intervention when they start. This may also be addressed by rolling-recruitment into the trial.
- The process of hiring may be different from organisation to organisation. This may cause large differences between clusters that may affect the overall results.

Manipulation checks

To measure whether we have addressed the trial threats, and test that our intervention is effective we would measure how gendered the language is in each job ad written. We would expect that language is *less* gendered in job ads written by the treatment group, compared with the control group.

Randomising job titles - cyber security focus/ICT focus

Overview

In this trial we would only change the *job titles* in real job ads, to test if this changes the proportion of women applying for jobs. We would change the titles to be either more masculine, or feminine stereotyped. For example, "cyber security analyst", to "audit-cyber security". We expect this to have an effect on the gender balance of applicants as the job title is the most salient part of the job ad when searching on online job boards.

We would need to partner with several organisations to randomly update the job titles in their job ads. We may need a large number of organisations to get a sufficient number of job titles.

Table 3. Pros and cons of a trial of only changing job titles

Pros	Cons
• Easy to implement intervention. • Testing whether the job title has an effect on whether people apply. • Reframing job title could be applied in a number of sectors.	• We may have to recruit a large number of organisations to achieve a powered sample. • It may be a time intensive intervention to change the title for a large number of job ads. • The effect size may be small or a null, there may be more effective intervention points. • We might not be able to convince organisation to let us change the job title. • Participating organisations may 'learn' the new job titles and start using them in business as usual job ads.

Trial aim

To see the impact of changing job titles on the demographics of who applies for the role.

Trial design

This would be a randomised controlled trial, where we randomly allocate job ads to have different job titles (without changing anything else), as illustrated in Figure 7. We would need to sign up multiple organisations in order to get a sufficient sample size. Participating organisations would submit job ads to BETA, and we would randomly change the job title.

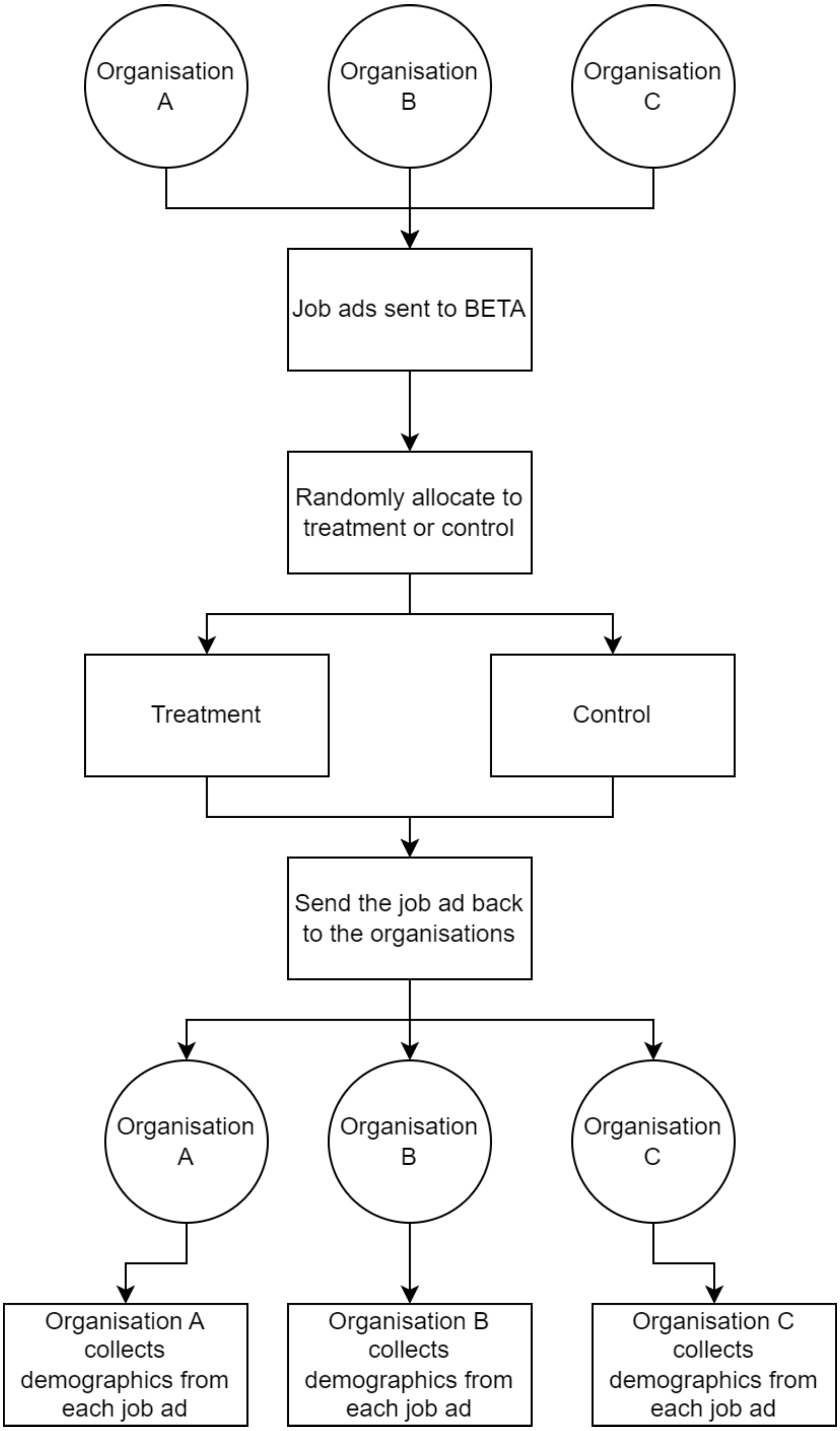
Organisation A
Organisation B
Organisation C
Job ads sent to BETA
Randomly allocate to treatment or control
Treatment
Control
Send the job ad back to the organisations
Organisation A
Organisation B
Organisation C
Organisation A collects demographics from each job ad
Organisation B collects demographics from each job ad
Organisation C collects demographics from each job ad

Figure 7. Diagram of the altered job titles trial design.

Intervention

Based on Gaucher et al. (2012) and BETA's machine learning analysis (BETA 2023) we would create a list of equivalent but less masculine job titles for cyber security/ICT roles. This list would have a corresponding masculine and feminine job titles, which we would use to update the ads.

The control job ads would have the business as usual job title.

Outcome measures

Primary outcome

The primary outcome is the proportion of male applicants who applied for each ad. To measure this, we would collect demographic information of applicants in the first round of applying (e.g. at the CV submission/cover letter).

Secondary outcomes

There are a number of secondary outcomes we could consider:

- Other demographics collected at the initial stage of the trial, e.g. age, and CALD, indigenous and disability status. Given the language is more inclusive we would expect that there may be crossover effects of the intervention.
- The demographics of the people successfully employed.
- The number of applicants in each group.

We would also collect the characteristics of each job ad (level of seniority, pay) to include as covariates in the analysis.

Data collection

We would rely on each partner organisation to collect demographics of applicants, at the point that applicants submit their CV.

Population and sample selection

We would need to partner with multiple companies. Our sample would be the job ads from organisations that agreed to work with us. This would result in selection bias, and reduce the generalisability of the results.

Hypothesis

The proportion of *male applicants* would be **lower** for job ads which had *less* stereotypically masculine job titles compared with control.

Randomisation

We would conduct randomisation at the job ad level, stratified by the organisation that we signed up.

Sample size and power

As mentioned, Burn et al. showed that including 3 ageist machine learning phrases reduced the proportion of older applicants by 12 percentage points, changing the proportion of older from 19% to 31%. We would only be changing the job title so would not expect to see as large an effect. We also may be constrained in how much we can change the job title: we can't change it so much it doesn't describe the job. Therefore, we limit our maximum treatment effect to 6 percentage points. We also don’t consider the stratification in this power analysis, and therefore present a conservative estimate of the power.

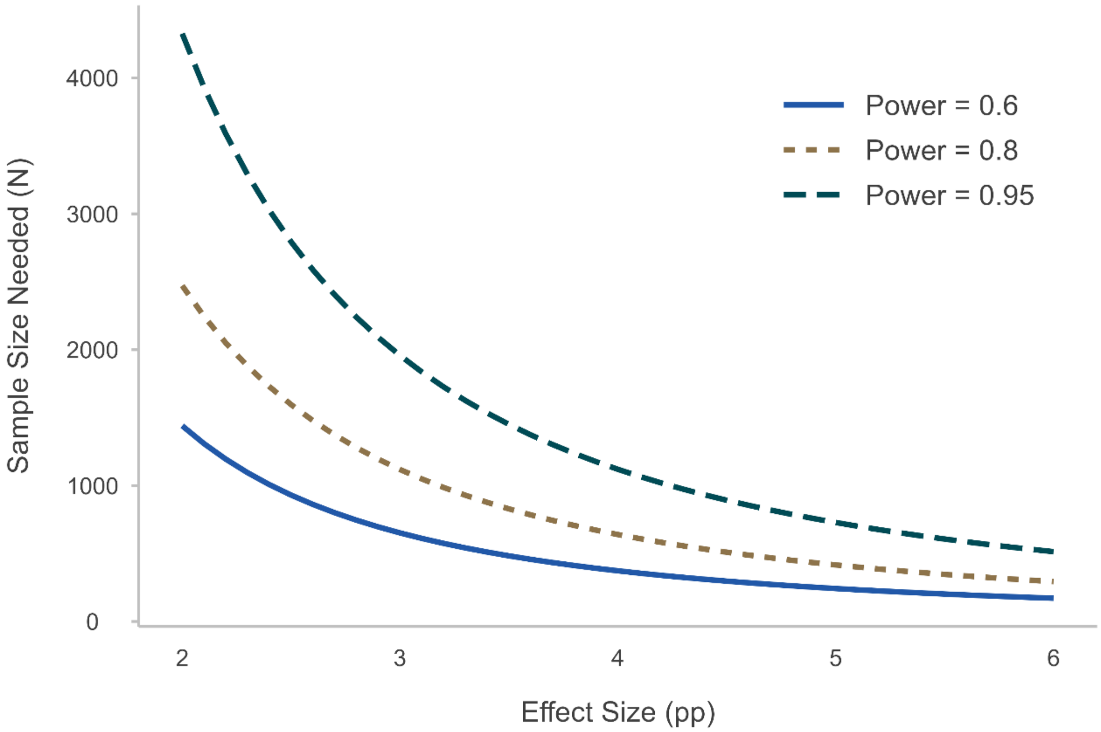

Figure 8. Power analysis varying the effect size and the outcome on the needed sample size.

References

BETA (2023). Attracting a diverse cyber security workforce: lessons from an analysis of Australian job ads. Report published at https://behaviouraleconomics.pmc.gov.au/sites/default/files/projects/attracting-diverse-cyber-security-workforce.pdf

Burn, I., Firoozi, D., Ladd, D., & Neumark, D. (2022). Help Really Wanted? The Impact of Age Stereotypes in Job Ads on Applications from Older Workers. *National Bureau of Economic Research Working Paper Series, No. 30287.* https://doi.org/10.3386/w30287

List, J. A. (2023). The Voltage Effect. *Business Economics*, *58*(1), 3-8.

Gaucher D, Friesen J, and Kay AC (2011) Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology,* 101(1):109-28. https://doi.org/10.1037/a0022530