Australian Government
Department of the Prime Minister and Cabinet

BEHAVIOURAL ECONOMICS TEAM
OF THE AUSTRALIAN GOVERNMENT

Unconscious Bias in Hiring in the Australian Public Service

Proposal for a Randomised Controlled Trial on Shortlisting

Michael J. Hiscox
BETA & Harvard University
29 June, 2016

SUMMARY

Background

Women are approximately 50% of the Australian workforce, but are under-represented in management and executive-level positions in the private sector and in many areas of the Australian Public Service (APS). In 2015, women comprised 58.4% of the APS as a whole, but accounted for only 42% of the Senior Executive-Level positions.[1] These statistics may reflect gender discrimination in hiring and promotion decisions. Such discrimination is generally difficult to overcome because cognitive biases can be so internalized that individuals are unaware that their decision making processes are affected. Even when hiring and promotion committees are briefed and trained to be attentive to gender equity and potential discrimination, implicit or unconscious bias can still play a large role. Creating gender-blind processes for reviewing job applications may provide an effective way to mitigate unconscious bias.

Objectives

The proposed study aims to assess the magnitude of gender bias in standard APS processes for selecting shortlists of job candidates and whether the introduction of a gender-blind approach to reviewing job applications can help eliminate bias. The primary outcomes we will measure are reviewers' initial assessment of whether a female applicant is 'potentially suitable' to the job or not, the rating scores (5-10 cardinal scale) reviewers give to females they initially select as 'potentially suitable' for the job, and

[1] Retrieved from the Australian Public Service Employment Database for the 2015 calendar year (Australian Public Service Commission, 2015).

whether any of the ‘potentially suitable’ female applicants are recommended among the top 5 applicants to be shortlisted for further consideration for appointment to a senior position in the APS.[2,3]

Design

The trial would be an individually randomised controlled trial, conducted in partnership with fifteen APS agencies. The trial would be a “framed field experiment” in which subjects drawn from senior and executive-level officers in these agencies are asked to review job applications for an executive-level position, a familiar task, but understand that this exercise is part of an experiment and their behaviour is being studied.[4] Subjects either review applications in the usual way based upon curriculum vitae or they review the same applications in de-identified (gender blind) form.

Costs and Timeline

For each participating APS agency, the principal costs associated with the study involve approximately 1-2 hours of time from each agency employee reviewing applications. The study would be conducted in October-November 2016. Findings would be reported in November-December 2016.

1. Background

Women are approximately 50% of the Australian workforce, but are under-represented in management and executive-level positions. This is evident not just in the private sector but also in many areas of the Australian Public Service. In 2015, women comprised 58.7% of the APS as a whole, but accounted for only 41.8% of the Senior Executive Service (SES) of the APS. In the highest two levels of the SES (Bands Two and Three), women hold only 36% of positions.[5]

These statistics reflect a range of factors affecting employment decisions by individuals and encompassing a variety of institutional and societal structures.[6] They also reflect gender discrimination in hiring and promotion. Such discrimination is generally difficult to overcome because cognitive biases can be so internalized that individuals are unaware that their decision-making processes are affected.[7] Even when hiring and promotion committees are briefed and trained to be attentive to gender equity

[2] The initial assessment of suitability will involve participants sifting through all 16 CVs and deciding whether a CV is ‘potentially suitable’ or ‘not suitable’ for the job. We will recommend that participants choose about 10 ‘potentially suitable’ CVs, which they will then rate on a more precise cardinal scale.

[3] A cardinal scale indicates the order of rankings, as well as the degree by which rankings differ. Under this particular cardinal scale, participants are allowed to give applicants any rating between 5-10 (with precision of up to 1 decimal place), with 5 being suitable for the position and 10 exceptionally suitable for the position. Additionally, participants must ensure that no two CVs receive the same score.

[4] For a detailed overview of the taxonomy of field experiments and how “framed field experiments” fit into this taxonomy, see Harrison & List (2004).

[5] Data retrieved from the Australian Public Service Employment Database for the 2015 calendar year (Australian Public Service Commission, 2015).

[6] One particular issue is the availability and adoption of flexible work arrangements making it easier for men and women to balance demands of work and family. Such arrangements are available in APS agencies, but are difficult to access and are accessed overwhelmingly by women and rarely if ever by individuals in senior management positions. See Australian Public Service Commission (2016a).

[7] Unconscious discrimination as an alternative to models of conscious discrimination (e.g. statistical and taste-based discrimination), requires thinking about drastically different solutions to overcome discrimination, and emphasise methods that do not force individuals to make decisions against their will (Bertrand, Chugh, & Mullainathan, 2005).

and potential discrimination, implicit or unconscious bias can still play a large role and weigh against female candidates for management positions. Recruiters and reviewers appear to be prone to a form of "homophily," favouring candidates with similar characteristics to their own.[8] There is also considerable evidence of large biases in hiring processes associated with female and minority identity characteristics.[9, 10]

One potentially effective approach to overcoming implicit bias may be to alter the standard procedure for reviewing and shortlisting candidates for management positions, in which the gender of each candidate is identified and known by reviewers, and instead adopting a new "gender-blind" procedure in which identifying information is removed from application materials. This approach is modelled on reforms adopted in the 1970s and 1980s by American symphony orchestras aimed at overcoming biases in hiring by introducing a screen during auditions to conceal the identity of the musician from the jury evaluating the performance. In a well-known study analysing data on auditions and hiring by orchestras over this period, Goldin & Rouse (2000) found that the use of blind auditions had a major impact on gender bias in orchestras, increasing the likelihood of female musicians being selected by 25-40%.[11]

Creating de-identified (and hence gender-blind) processes for reviewing job applications may provide an effective way to mitigate unconscious bias. Anonymous job application procedures have been introduced in small-scale ways and at different levels of government in Sweden, the Netherlands, Belgium, and Switzerland, and implementing legislation has been proposed in Britain and has actually passed in France.[12] In Australia, in May 2016, the Australian Bureau of Statistics (ABS) reported that it had implemented an anonymized application process during a recent hiring round and the Victorian government announced that it would be trialling a similar process for recruitment for several types of government positions in 2017.[13]

[8] Results across a range of studies suggest that individuals use questionable arguments to justify implicit biases around gender and race (Norton, Vandello, & Darley, 2004).

[9] A path-breaking study in 2004 measured racial discrimination in the United States job market by sending fictitious resumes to help-wanted ads in newspapers, randomly assigning to the resumes African-American- or white-sounding names. White names received 50 percent more callbacks for interviews (see Bertrand & Mullainathan, 2004). This study has been emulated in Europe and Australia: see Carlsson & Rooth (2007); Kaas & Manger (2012); Booth, Leigh, & Varganova (2012).

[10] Gender-based correspondence studies point to bias against women being particularly pronounced in high-status or male-dominated jobs, while there is also evidence that discrimination against men is present in jobs that are dominated by females. Moss-Racusin, Dovidio, Graham, & Handelsman (2012) conduct a double-blind correspondence study to test gender bias of University faculty in hiring students to a laboratory manager position, finding that both female and male faculty were biased towards selecting male applicants. Heilman, Wallen, Fuchs, & Tamkins (2004) conduct a series of lab experiments that employ hypothetical CVs to test subjects' reactions to a woman's success in jobs strongly associated with males; they find that unconscious bias can lead to competent women being disliked, which subsequently has detrimental effects on the probability of receiving a salary increase. For an overview of correspondence studies in gender bias, see Azmat & Petrongolo (2014) and Bertrand & Duflo (2016).

[11] To identify the effect of adopting the blind auditions, Goldin & Rouse (2000) use a differences-in-differences design, relying on an assumption that the introduction of the new process is not simultaneous with any other changes that affected diversity in the orchestras. One possibility is that the adoption of the blind auditions actually encouraged larger numbers of talented female musicians to apply for positions by signalling that the orchestras were serious about addressing discrimination – a self-selection effect that was separate from the effect on reviewer choice.

[12] See Krause, Rinne, & Zimmermann (2012).

[13] For news reports on the ABS and Victoria government initiatives, see respectively Towell (2016) and Dean (2016).

Several studies conducted in public and private sector contexts in European countries in recent years suggest that de-identification of job applications could significantly reduce bias. One study examined the effects of the introduction of de-identified job application forms for a group of public service positions in Sweden in 2006. Job applicants were required to follow a specific anonymous application procedure, which was highlighted in the job advertisements. The findings indicated that the new anonymous process equalized the probability of being interviewed and hired for male and female candidates.[14] Another, larger study was conducted in Germany in 2010 with participation from eight private and public sector organisations which agreed to alter their review processes and removed characteristics of applicants (including name and contact details, gender, nationality, date and place of birth, dates of previous employment, disability, marital status and the applicant's picture) from application materials before they were reviewed. Compared with previous job searches by these organisations, the data indicated that the switch to de-identified processes appeared to increase call-back rates for women and ethnic minorities in many, but not all, cases.[15]

The only large-scale field experiment to date, to our knowledge, was conducted in 2010-2011 by the French Public Employment Service (PES) and involved around 1,000 firms filling job positions in urban locations over a 10 month period.[16] Job positions posted at the PES by participating firms were randomly assigned to either the standard procedure or a new de-identified approach when the PES sent the firms the set of resumes of qualified job-seekers who apply or have previously been registered at the PES. For the de-identified approach, research assistants erased the first part of each resume containing any information about the applicant's name, address, gender, age, marital status, number of children, and any photos. In both treatment and control conditions the firms then selected short lists of applicants for interviews. The findings indicated that, when compared with the standard process, the de-identified process appeared to improve the chances of women being selected for interviews, but also appeared to reduce the probability of call backs for interviews among ethnic minorities. The study found no evidence that the anonymous review process affected the cost of hiring for firms in terms of the time and number of interviews needed to fill a position.[17]

Collectively these existing studies suggest that de-identifying applications can result in higher initial call-back rates for interviews for women and perhaps also for ethnic minorities if discrimination is present initially, however de-identification may have no impact in contexts in which no discrimination is present

[14] Åslund & Skans (2012) use a differences-in-differences approach similar to that used by Goldin & Rouse (2000), and thus has similar limitations in terms of potential time-varying confounders and the potential self-selection effects on the composition of applicants.

[15] Krause, Rinne, & Zimmerman (2012) note that the study is limited by relying on previous job searches to provide the "control" group or counterfactual to assess the impact of the de-identification treatment.

[16] See Behaghel, Crépon, & Le Barbanchon (2012).

[17] The authors are careful to point out a number of limitations of the experiment. In particular, participating firms knew that they were part of an experiment addressing discrimination and this may have affected their behaviour in the control condition in which identities of applicants were known. If firms are concerned not to be seen to be discriminating against minorities, this might account for why firms in the control condition appear to make choices that are more favourable for minorities than firms using the anonymized process. It also suggests that the measure of the impact of de-identification on gender bias may be attenuated. The authors also note that the de-identification of resumes was not comprehensive. In particular, the gender identity of applicants may have been inferred from specific terms used and information provided (e.g., schools attended) in the body of each resume.

initially and can also undermine existing affirmative action measures.[18] Behaghel, Crépon, & Barbanchon (2015) provide evidence that employers who tended to hire minority candidates at higher rates were also more likely to volunteer to participate in their changed recruitment trial, highlighting the importance of being cautious in generalising the findings of field experiments in discrimination where participation is voluntary, as there may be strong self-selection effects.

The findings to date are somewhat mixed, and these studies are limited in many ways in terms of the methodologies applied. If unconscious bias operates most strongly at later stages of the recruitment process, there is also a concern that de-identifying applications will have little or no impact in decreasing unconscious bias in recruitment. This is a reasonable concern since most recruitment practices employ interviews at later stages of the process. Nonetheless, the full impacts of introducing de-identified application processes are still largely unknown, and this is key step in enhancing our understanding how unconscious bias in recruitment operates.

2. Objectives

The proposed randomised controlled trial (RCT) aims to assess the magnitude of gender bias in standard Australian Public Service (APS) processes for selecting shortlists of executive-level job candidates and whether the introduction of a gender-blind approach to reviewing job applications can help eliminate any existing bias. In particular we will address whether de-identification of resumes of job candidates improves the assessments of female candidates and their chances of being selected for shortlists and included on ranked lists for job offers.

The primary outcomes we will measure are reviewers' initial assessment of whether a female applicant is 'potentially suitable' to the job or not, the rating scores (5-10 cardinal scale) reviewers give to females they initially select as 'potentially suitable' for the job, and whether any of the 'potentially suitable' female applicants are recommended among the top 5 applicants to be shortlisted for further consideration for appointment to a senior position in the APS. We will also examine the costs of the hiring process in terms of the time required by reviewers to assign ratings and select short lists and ranked lists for job offers, comparing the time spent between the control and treatment groups. Additionally, we will also examine pure gender differences between male and female names, as well as the effects of ethnic discrimination. The full set of primary and secondary outcomes are explained in detail in **Appendix 3 and Appendix 4.**

3. Design

a. Partners and Subject Pools

The trial will be conducted in partnership with fifteen agencies within the Australian Government: the Australian Public Service Commission (coordination), the Australian Tax Office, the Treasury, the Department of Employment, the Department Social Services, the Department of Defence, the Department of Health, the Department of Environment, the Department of Industry, Innovation and Science, the Department of Foreign Affairs and Trade, the Department of Agriculture, the Attorney General's Department, the Fair work Ombudsman, the Office of National Assessments, and the Depart of the Prime Minister and Cabinet.

[18] See also Bøg & Kranendonk (2011); Krause, Rinne, & Zimmermann (2012); Behaghel, Crépon, & Le Barbanchon (2015).

Individuals eligible to participate as subjects in the trial are all Executive Level (EL) and Senior Executive Service (SES) Band 1 officers within the participating agencies. These are individuals who would normally be involved in recruitment and hiring decisions for EL positions within their own agencies. Across all fifteen participating agencies, the total population of eligible subjects is approximately 22,000 individuals.

b. Randomisation of the Intervention

The trial would be an individually randomised controlled trial. Specifically it would be a "framed field experiment" in which subjects are drawn from the population of interest and complete a familiar task, something they would normally perform in a naturally-occurring context, but understand that in this instance they are completing the task as part of an experiment and their behaviour is being studied.[19]

Eligible individuals will be invited to participate in the study via an e-mail message sent by the head of corporate for their agency (see **Appendix 1**). The invitation will request that each officer participate in the study to help improve recruitment practices within the APS, and explain that the task will involve reviewing 16 hypothetical candidates and selecting a shortlist for a hypothetical appointment to an EL2 position in their agency, but provides no details of the specific issues being addressed by the research or any mention of de-identification of applications. Individuals register to participate in the study by clicking on a link embedded in the message and are notified of the date on which they will be sent materials to review and the rating scheme to be used by all reviewers.[20]

Individuals who agree to participate in the study and register online will then be randomly assigned to review materials in either the de-identified (treatment) or usual (control) condition, with pre-randomisation stratification based upon agency, age, gender, and APS level. Additionally, we will have **two control groups and one treatment group** such that the only difference between the two control groups is that the first names used for the CVs in control group 1 will be swapped to similar first names of the opposite gender (e.g. the name Jane Smith in control group 1 would be John Smith in control group 2). Having these two control groups would allow us to conduct more detailed ethnicity tests.

On a designated date participants will be sent via email a link to an online survey form in which they can view candidate materials, select who they think are 'potentially suitable' candidates, and subsequently assign rating scores (on a 5-10 cardinal scale) to their chosen 'potentially suitable' candidates. At the end of the form they will be asked to select the top 5 candidates for the shortlist for more serious consideration for the position. Each individual subject will be asked to complete and submit the online form within 5 working days of receiving it.

The candidate material will consist of a 2-page curriculum vitae. The research team will generate CVs for 16 hypothetical candidates. To prepare the application materials for our hypothetical candidates, we first standardize the format in which information in each curriculum vitae is presented. We allow no variation in font size, type, or colour and no other graphical features (e.g., photographs).

[19] Harrison & List (2004).

[20] A similar approach is used to collect responses for the annual APS employee census, which is sent to all APS employees. Employees are sent an email with a link to complete the census, and they understand their responses are confidential, and that data is only used in an aggregate, de-identified form (Australian Public Service Commission, 2016b).

For the curriculum vitae, the information is presented in the following order:

a. Name, Address, Telephone Number and E-mail Address (for the identified versions of the CVs)
b. Personal details: Security clearance level, citizenship status, Indigenous status (for identified versions of the CVs), ongoing disability (for identified versions of the CVs).
c. Objective: the position sought and match to the person's career experience and objectives
d. Key skills: list of specialized skills relevant to position sought
e. Education: list by degree, major, institution attended, graduation date
f. Experience: list by position, organisation, dates of employment, and responsibilities
g. Additional information: list of language and computer skills, and special interests.
h. References: "References available upon request"

Standardizing the information from curriculum vitae in this way represents a departure from the "business-as-usual" process which allows candidates to add more material and make stylistic and presentational choices. We choose to standardize material in this way in order to create de-identified application material in the most efficient and least-distracting way – by simply removing section (a). This standardization is also the form in which de-identification could be implemented at scale in hiring processes, subsequently, via dedicated software platforms.

In addition, standardization helps to minimize the amount of "noise" created in evaluations by stylistic and presentational choices and we do not have to be concerned that reviewers may infer gender from stylistic or presentational choices (e.g., font, formatting). Another advantage of standardization of materials is that, subjects tempted to think about the purpose of the research may infer that the study is assessing the effects of creating standardized application materials, rather than guessing that the research is addressing gender discrimination, so standardization should help to minimize the risk of subject reactivity in the form of more abnormally favourable treatment of female candidates in the control (identified) condition.

The materials describe a set of 16 realistic candidates who vary in terms of educational institutions (top 8 Australian universities vs other Australian universities), postgraduate qualifications (postgraduate degree vs only bachelor degree), and work experience (years of work experience and type of work experience). While attempting to make the 16 CVs as realistic as possible, we also try to ensure that the CVs have a similar overall level of quality.

We pre-test the CVs among a sample of about 30 staff members across our Department who are knowledgeable on recruitment practices in the Australian Government, as well as with about 15 SES-level volunteers from agencies participating in the trial to ensure that the online security system of each participating agency allows staff to receive external emails with instructions to participate in the exercise. Pre-testing ensures that CVs have a similar overall level of quality, and that there is good variation in average evaluation scores across these candidates and overall gender balance in terms of the expected evaluation scores and expected proportion of men and women in the top-5 of the ranked lists.[21] We distribute different quality characteristics evenly across individuals to maximize variation in ratings and selections across subjects and allow room for unconscious bias to operate – in particular, we aim to avoid an outcome in which all or most subjects converge on the same top 5 list to shortlist as the

[21] We also pre-test the names on the CVs with staff in our Department to ensure that they have a strong association with a given gender, in order to avoid the possibility that participants may incorrectly infer the gender of an applicant.

obvious best picks; this is an important precaution given evidence that the extent to which unconscious bias operates is context-dependent.[22]

Figure 1 shows the schedule for participant progress through the trial as well as data collection.

c. Sample Size and Statistical Power

A total sample of approximately 1,400 will provide at least 80% power for two-sided tests (at the 5% significance level) to detect substantial (≥5%) differences in outcomes between the treatment and control groups for the primary outcomes related to de-identification. **Table 2** shows the power calculations for primary outcomes using the results of our pre-testing exercise (See **Appendix 2** for more detailed results of calculations showing power and minimum detectable effect sizes for different samples).

Table 2: Primary Outcomes and Statistical Power for Samples of 800-1400

Outcomes	Usual Practice mean	Anticipated Effect %	Anticipated De-Identified mean	Power: N=800	N=1000	N=1400
Women who are 'potentially suitable' for job (%)	43.0	5%	45.15	0.99	0.99	0.99
Women on shortlist (%)	43.0	5%	45.15	0.99	0.99	0.99
Women Avg. Score (on a cardinal scale of 5- 10)	7	5%	7.35	0.99	0.99	0.99
Women Avg. shortlist ranking (on an ordinal scale of 1-5)	3	5%	2.75	0.59	0.69	0.83

Notes: Data from small-sample pre-testing of CVs among PM&C volunteer staff.

In particular, a sample of 1,400 provides 99% power for a two-sided test (with 2.5% alpha level) to detect a 5% change the percentage of women rates as 'potentially suitable' for the job or shortlisted (from 43% to 45.2%). It also provides 99% power to detect a 5% change in the average evaluation scores for female candidates (from 7 to 7.35 out of 10), and 83% power to detect a 5% change in the average shortlist rankings of women (on a 1-5 ordinal scale). Notice that the alpha level for these tests for effects for each outcome is set to 2.5% so that the overall Type One error rate for testing for statistical significance of effects is kept at the standard 5% level.

Given the available baseline information, the power calculations for the pure gender effects are qualitatively the same as described above for the effects of de-identification, so we have omitted them. There is no baseline information on the primary outcomes for ethnicity, so we cannot generate power calculations for these outcomes (see **Appendix 3 and Appendix 4** for a detailed discussion of our outcomes of interest). It is possible that since the magnitude of some of the outcomes for ethnic minorities is smaller, some of our ethnicity tests may be underpowered.

[22] Conditions where unconscious bias operate strongly include scenarios where individuals are distracted or face tight time pressures, rely on vague criteria, undergo certain emotional states (e.g. anger), perceive social categories such as race or gender as salient, face little accountability for their decisions, or complete tasks that require little cognitive effort (Casey, Warren, Cheesman II, & Elek, 2012).

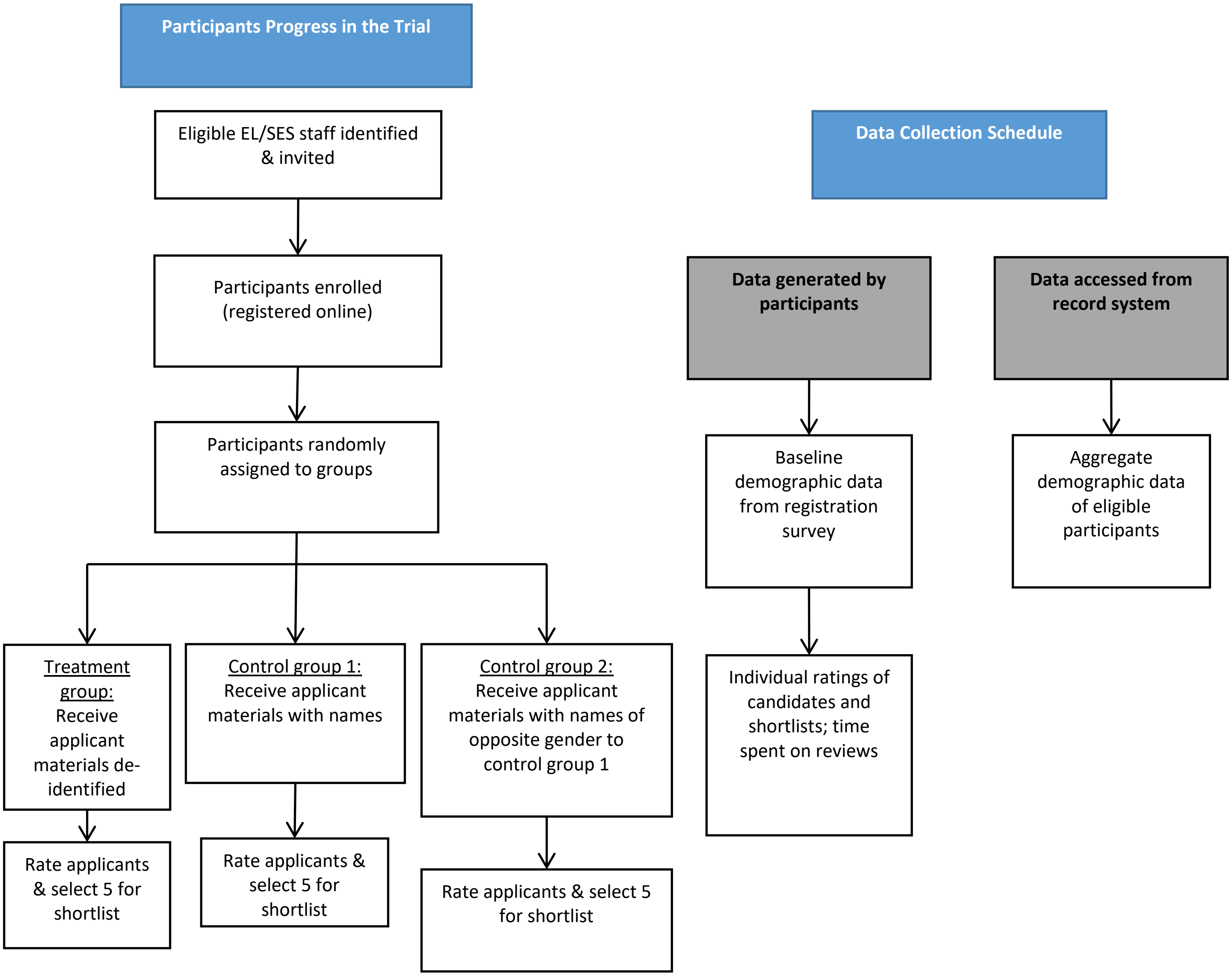

Participants Progress in the Trial
Eligible EL/SES staff identified & invited
Participants enrolled (registered online)
Participants randomly assigned to groups
Treatment group: Receive applicant materials de-identified
Control group 1: Receive applicant materials with names
Control group 2: Receive applicant materials with names of opposite gender to control group 1
Rate applicants & select 5 for shortlist
Rate applicants & select 5 for shortlist
Rate applicants & select 5 for shortlist
Data Collection Schedule
Data generated by participants
Data accessed from record system
Baseline demographic data from registration survey
Aggregate demographic data of eligible participants
Individual ratings of candidates and shortlists; time spent on reviews

Figure 1 Participant Progress in the Trial and Schedule for Data Collection

d. Data

Baseline data on subjects will be gathered from participants as they register to participate in the trial. Participants will be given a week to register for the trial. The key baseline data on subject characteristics will include age, gender, APS level, experience with EL-level hiring rounds, and partner agency. These characteristics will be used to stratify the sample prior to random assignment to treatment and control groups. This stratification will enhance the statistical power of the study and also allow for comparisons of the effects of de-identification across key sub-groups (e.g., male vs female reviewers, younger vs older reviewers).

Once subjects receive candidate materials and complete the online review process, we will collect data on the number of women that participants initially rate as 'potentially suitable' for the job and on the rating scores (5-10 cardinal scale) reviewers give to females they initially select as 'potentially suitable' for the job. We also collect the list of 5 candidates selected by each subject for the short list, and the software will record the total time taken by each subject to review the materials and complete the process of rating and shortlisting.

e. Analysis Plan

Balance checks for the treatment and control groups will be conducted and reported for all key individual-level covariates measured at baseline, including participants' age, gender, APS agency, APS level and type of role, and their self-reported previous experience with EL-level hiring rounds.

The critical question is whether de-identification of applications of job candidates improves the relative assessments of female candidates and their chances of being selected for shortlists. There are three possibilities. If there is a bias against female candidates in the usual setting, de-identifying applications should cause an increase in average ratings for women and raise the chances of them being selected for shortlists. If there is no bias in the usual setting, de-identification of applications should have no impacts. If reviewers actually engage in a form of affirmative action in favour of female candidates as part of usual practice, de-identification should actually decrease average rating for women and their chances of selection onto shortlists.

We will also conduct a simple test to examine the effectiveness of the de-identification procedure. We will analyse the data on ratings and shortlist selections from subjects in the treatment group only and test whether ratings and the probability of shortlist selection vary systematically by candidate gender. If de-identification is effective there should be no such systematic relationship. This should be the case if we have created gender balance in the way we have distributed educational and other characteristics among hypothetical candidates, so that subjects cannot guess the gender of the candidates based on de-identified materials.

The principal analysis of the effects of the intervention will be comparisons of primary outcomes (the percentage of women listed as 'potentially suitable' for the job, the ratings (5-10 cardinal scale) for female candidates listed as 'potentially suitable' for the job, and the percentage of women recommended for shortlists) across the treatment and control groups to estimate average treatment effects. Confidence intervals for these estimates (and *p* values) will be based upon regression analysis that includes the treatment indicator along with all covariates used in pre-randomisation stratification (See **Appendix 3 and Appendix 4** for more details on the analysis plan).

Secondary analysis will compare outcomes across treatment and control groups for pre-specified sub-groups (e.g., based upon gender, age, APS level and role type, and agency) to explore potential heterogeneity in treatment effects. Given preliminary evidence that unconscious bias has heterogeneous effects across gender and ethnic minorities as well as that less than 16% of APS employees identify as being from a Non-English speaking background, we will nest a test of ethnicity within this trial.[23,24] In order to nest this test of ethnicity we will use a portion of names that are strongly associated with ethnic minorities. We will also have two control groups as detailed previously, and would thereby be able to test how ethnic bias effects vary depending on the gender of a particular

[23] For evidence of heterogeneous effects, see Behaghel, Crépon, & Le Barbanchon (2012).

[24] Data on the number of APS employees from a Non-English speaking background was retrieved from the Australian Public Service Employment Database for the 2015 calendar year (Australian Public Service Commission, 2015).

minority. Including names strongly associated with ethnic minorities would also make the exercise more realistic given that a large proportion of Australia's population has been born overseas.[25] While throughout this proposal we refer to testing gender bias, we will also test the effect of ethnicity bias across primary and secondary outcomes analogous to those discussed for gender bias. Since the trial is designed with power only sufficient to detect sizeable overall differences in average outcomes between treatment and control groups, secondary analysis of sub-groups may be somewhat underpowered in specific cases.

Given the recent focus on gender equality and minimising unconscious bias in the APS, there is a possibility that participants in the control groups will suspect that the study is testing gender bias and rate females more generously than they otherwise would. To guard against this potential drawback, we have included post-survey questions that will help us identify participants who may have become aware of the trial's focus.

f. Ethics

The study proposal will be submitted for review and approval by the Committee on the Use of Human Subjects in Research (Institutional Review Board) at Harvard University. Since we are not using real job candidates, participants understand that their behaviour is being studied, and all information about individual participants in the study (including review scores assigned by each participant) will be kept confidential, we expect expedited ethics review and approval.

[25] Booth et al. (2012) conduct a field experiment to measure ethnic discrimination in Australia for entry-level jobs using a correspondence methodology, and find that ethnic minorities suffer from discrimination, but that the degree of discrimination varies systematically across ethnic groups.

4. References

Åslund, O., & Skans, O. N. (2012). Do anonymous job application procedures level the playing field? *Industrial & labor relations review, 65*(1), 82-107.

Australian Public Service Commission. (2015). APS Employment Database. Retrieved from: http://www.apsc.gov.au/about-the-apsc/commission-services/apsed/apsedii

Australian Public Service Commission. (2016a). Balancing the Future: Australian Public Service Gender Equality Strategy 2016–19. Canberra.

Australian Public Service Commission. (2016b). 2016 APS employee census: Frequently Asked Questions. Retrieved from http://stateoftheservice.apsc.gov.au/2016-census-faqs/

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

Behaghel, L., Crépon, B., & Barbanchon, T. L. (2015). Unintended Effects of anonymous resumes. *American Economic Journal: Applied Economics, 7*(3), 1-27.

Behaghel, L., Crépon, B., & Le Barbanchon, T. (2012). Do anonymous resumes make the battlefield more even? evidence from a randomised field experiment. *IZA Journal of European Labor Studies, 1*(5), 1-20.

Bertrand, M., & Duflo, E. (2016). Field experiments on discrimination (NBER Working Paper No. w22014)

(pp. 1-110): National Bureau of Economic Research.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991-1013.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *The American Economic Review, 95*(2), 94-98.

Bøg, M., & Kranendonk, E. (2011). Labor market discrimination of minorities? yes, but not in job offers *MPRA Paper 33332*.

Booth, A. L., Leigh, A., & Varganova, E. (2012). Does ethnic discrimination vary across minority groups? Evidence from a field experiment. *Oxford Bulletin of Economics and Statistics, 74*(4), 547-573.

Carlsson, M., & Rooth, D. (2007). Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics, 14*(4), 716-729.

Casey, P. M., Warren, R. K., Cheesman II, F. L., & Elek, J. K. (2012). Helping Courts Address Implicit Bias. Williamsburg, VA: National Center for State Courts.

Dean, J. (2016). 'Blind' job applications: Victoria Government seeks to remove unconscious bias from employment. *ABC News*. Retrieved from abc.net.au/news/

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The Impact of "Blind" Auditions on Female Musicians. *The American Economic Review, 90*(4), 715-741.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic literature, 42*(4), 1009-1055.

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology, 89*(3), 416.

Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labour market: a field experiment. *German Economic Review, 13*(1), 1-20.

Krause, A., Rinne, U., & Zimmermann, K. F. (2012). Anonymous job applications in Europe. *IZA Journal of European Labor Studies, 1*(5), 1-20.

Krause, A., Rinne, U., & Zimmermann, K. F. (2012). Anonymous job applications of fresh Ph. D. economists. *Economics Letters, 117*(2), 441-444.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474-16479.

Towell, N. (2016). Public service goes blind to solve its women problem. *The Canberra Times*. Retrieved from http://www.canberratimes.com.au

Appendix 1: Invitation covering email and Invitation letter

UNCONSCIOUS BIAS IN SHORTLISTING TRIAL
DRAFT E-MAIL INVITATION TO PARTICIPATE

[You may wish to make annotations to this template email to suit the communication style of your agency and/or put the request in your voice. However, please do not alter the substantive content, instructions, or description of the research topic in the email. Please pass on to BETA any changes to the email that you intend to make before sending it out.]

Dear colleagues

Dear colleagues

The Australian Public Service Commission has invited our agency to participate in a trial to study recruitment processes in the Australian Public Service with the aim of developing new-and-improved approaches. I believe your input to this study will be valuable, and I have nominated you to participate.

Further details on the trial and invitation to participate are attached.

To register your participation in the trial, simply click on this link:

https://youropinion.au1.qualtrics.com/SE/?SID=SV_5msof7DgwG0Ib41

The exercise will require approximately one hour of your time between 14-18 November. Please complete the task in normal office hours, or whenever is most convenient for you.

There are several agencies participating in this study and I want our agency to lead by example. This is a great opportunity to demonstrate that we are committed to APS recruitment practices that hire the right people for the right jobs at the right time.

I want to thank you in advance for your participation and commitment to improving the APS.

Regards

[Head of Corporate]

Australian Government

Australian Public Service
Commission

We need your help to improve our recruitment process and build a better Australian Public Service. Please register today.

The Australian Public Service Commission is leading a study to evaluate a new approach to collecting and reviewing information on candidates in Australian Public Service recruitment processes. It focuses on the submission and shortlisting stages, and uses a digital platform that makes it easier to collect, distribute and assess relevant information.

The aim of this new approach is to help develop selection processes that are efficient and consistent, and put in place measures that uphold and promote the APS Values and Employment Principles relevant to recruitment and selection.

Feedback and advice from experienced public servants is being sought to help assess whether the new approach makes recruitment more systematic, consistent and efficient.

The exercise will require approximately one hour of your time during the week of 14-18 November.

To register your participation in the study, click on this link:

https://youropinion.au1.qualtrics.com/SE/?SID=SV_5msof7DgwG0Ib41

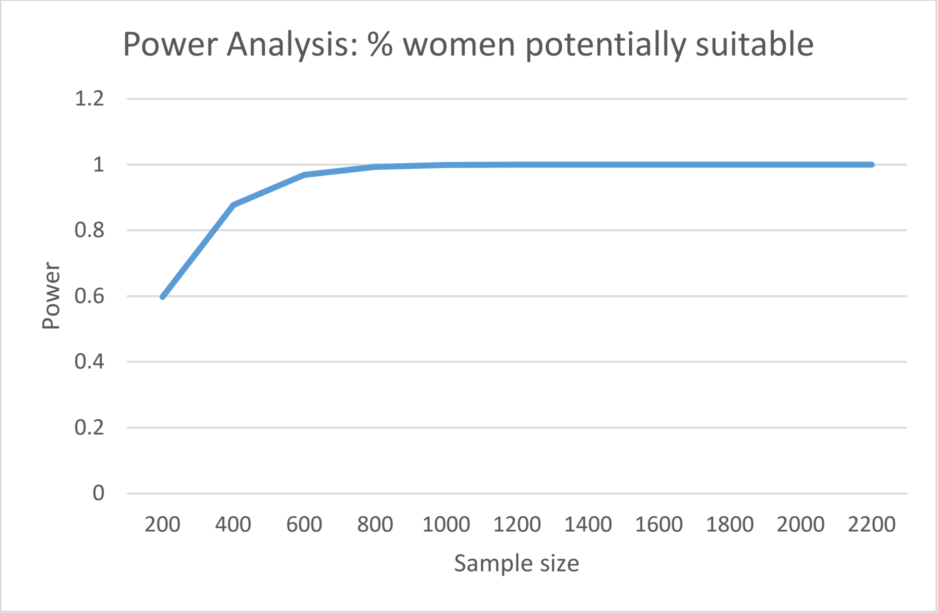
If you would like to participate, we recommend that you block out an hour in your calendar now for the week of 14-18 November to complete the exercise.

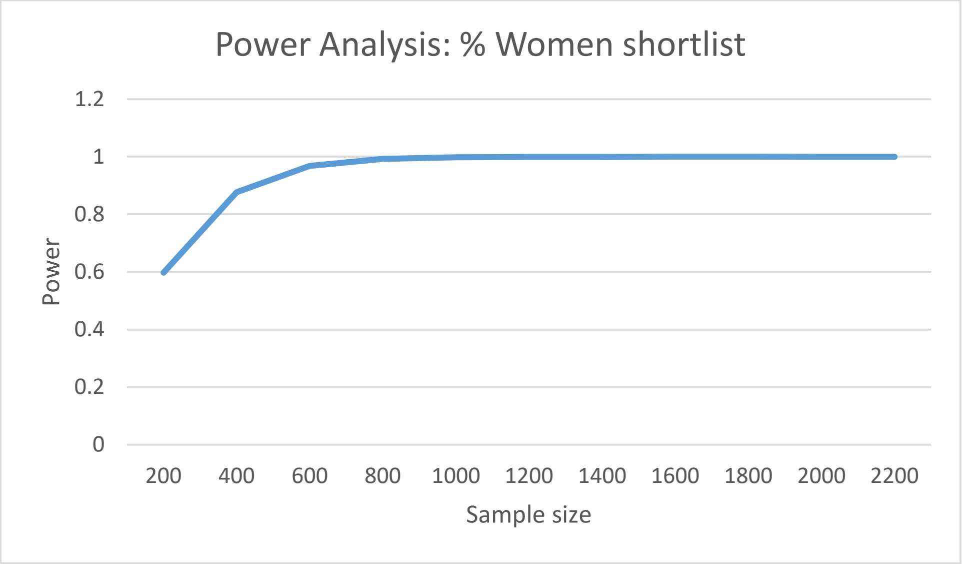
You will be supporting an important initiative. The results will be used to improve hiring practices across the APS and a report will be distributed to participating agencies.

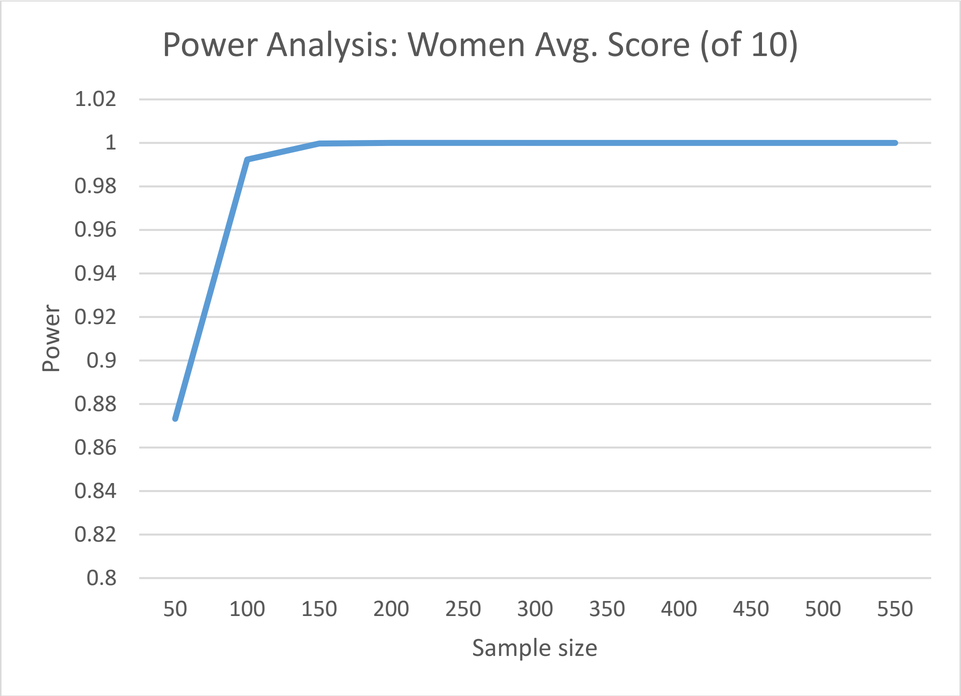
All of your responses will be kept confidential and unidentifiable.

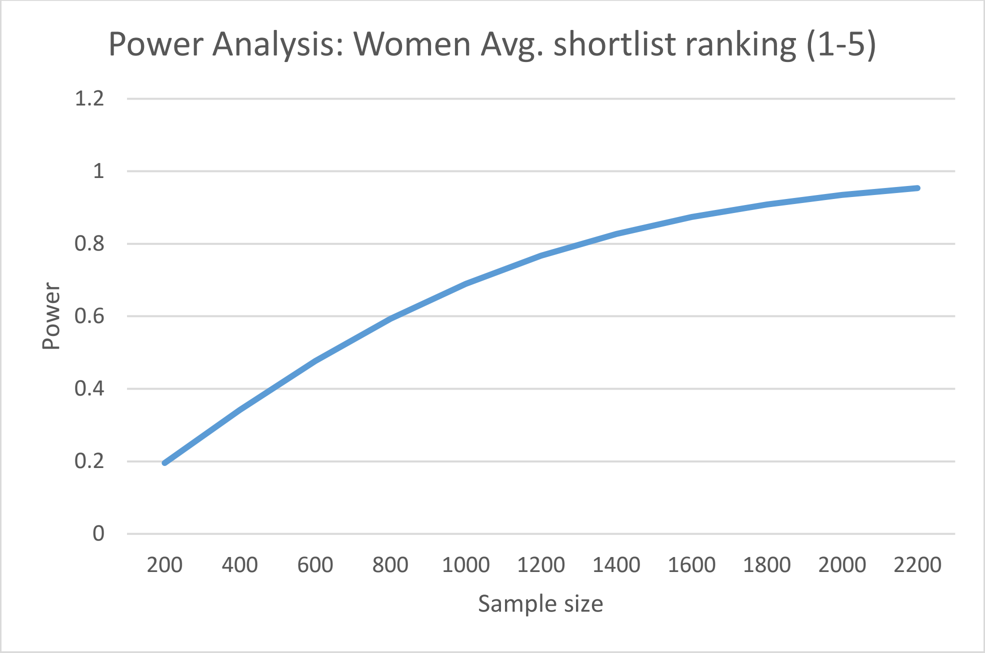
If you have any questions, you can contact the APSC at trial@apsc.gov.au.

Appendix 2: Power Calculations






Power Analysis: Women Avg. Score (of 10)
1.02
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
Power
50
100
150
200
250
300
350
400
450
500
550
Sample size


Power Analysis: Women Avg. shortlist ranking (1-5)
1.2
1
0.8
0.6
0.4
0.2
0
Power
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
Sample size

Appendix 3: Modelling the key outcomes of interest

Primary Outcomes of interest

1. Pure gender effects:

To purely capture the effect of gender discrimination, we will estimate regression models that compare the outcomes across the two control groups. For n_{c1} reviewers of control group 1 and n_{c2} reviewers of control group 2, we will estimate the following general model for a given outcome of interest Y_i for reviewer i:

$$Y_i = \beta_0 + \beta_1 C_i + \varepsilon_i$$

$$\text{where } i = 1, 2 \ldots, n_{c1}, n_{c1} + 1, n_{c1} + 2, \ldots\ldots, n_{c1} + n_{c2}$$

C_i is an indicator for whether a reviewer i is in the reference control group, and ε_i is the error term of the model. The unconditional average treatment effect is β_1 and this is the key parameter we are interested in. It is important to note that we will conduct this analysis using control group 1 as the reference group, and then control group 2 as the reference group. This will provide a quality check that the distribution of gender to CVs is the same in terms of quality across the control groups. The outcome measures we will analyse include:

- % CVs with female name in reference control group that are potentially suitable for job
- Average score (on a 5-10 cardinal scale) of potentially suitable CVs with female names in reference control group
- % CVs with female names in the reference control group that are shortlisted
- Average shortlist ranking (on a 1-5 ordinal scale) of CVs with female names in reference control group

Furthermore, as there is a possibility that gender discrimination differs between Anglo-Celtic names and names of ethnic minorities, we will conduct the above analysis in three different groups: for all 16 CVs; only for CVs with Anglo-Celtic names; only for CVs with minority names.

2. The effect of CV de-identification:

To capture the effect of de-identifying applications on gender discrimination, we will estimate regression models that separately compare the outcomes of females in control groups $C = 1, 2$, against the outcomes of the same CVs in the de-identified group. For n_c reviewers of control groups $C = 1, 2$, and n_T reviewers of the treatment group, we will estimate the following general model for a given outcome of interest Y_i for reviewer i:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

$$\text{where } i = 1, 2 \ldots, n_c, n_T + 1, n_T + 2, \ldots\ldots, n_c + n_T,\ C = 1, 2$$

The unconditional average treatment effect is β_1, and this is the key parameter we are interested in as it measures the effect of CV de-identification. The outcome measures we will analyse include:

- % CVs with female name in reference control group that are potentially suitable for job

- Average score (on a 5-10 cardinal scale) of potentially suitable CVs with female names in reference control group
- % CVs with female names in the reference control group that are shortlisted
- Time taken to assess candidate pool (minutes)
- Average shortlist ranking (on a 1-5 ordinal scale) of CVs with female names in reference control group

We will compare each control group (referred to as “reference group” above) to the treatment group in a separate model. Additionally, given that it is possible that de-identification could also have an effect on men, we will repeat the above analysis for the CVs with male names instead of female names.

3. Ethnicity analysis:
To capture the effect of ethnic discrimination, we will estimate regression models that compare the outcomes **within and between** the two control groups. The example below illustrates how we will analyse ethnic discrimination for one ethnicity, and we will conduct this analysis for each of the four ethnic minorities in the pool of CVs.

It is important to highlight that each gender/ethnic combination is fixed to a particular CV content. Although we have pre-tested the CV contents and tried to make them as similar as possible in terms of overall quality, it is still possible that on average reviewers may perceive a particular CV content as being weaker or stronger than the other CV contents. If this is the case, then we would incorrectly infer the effect of a weak/stronger CV content as an ethnic effect. To guard against this, we will test if a particular CV content is perceived as being weaker or stronger relative to the other CV contents on average, and qualify our ethnic analysis accordingly.

For the ethnic discrimination tests **between** control groups, we will be comparing how discrimination differs between the female and male candidate of a given ethnicity. Given n_{c1} reviewers of control group 1 and n_{c2} reviewers of control group 2, we will estimate the following general model for a given outcome of interest Y_i for reviewer i:

$$Y_i = \beta_0 + \beta_1 C_i + \varepsilon_i$$

$$\text{where } \ i = 1, 2 \ldots, n_{c1}, n_{c1} + 1, n_{c1} + 2, \ldots \ldots n_{c1} + n_{c2}$$

C_i is an indicator for whether a reviewer i is in control group 1, and ε_i is the error term of the model. The unconditional average treatment effect is β_1, and this is the key parameter we are after. The outcome measures we will analyse include:

- A dummy equal to 1 if the CV with the ethnic last name of interest is rated as potentially suitable for the job, and to zero otherwise.
- Score (on a 5-10 cardinal scale) if the CV with the ethnic last name of interest is potentially suitable for the job.
- A dummy equal to 1 if the CV with the ethnic last name of interest is shortlisted, and to zero otherwise.
- Shortlist ranking (on a 1-5 ordinal scale) of the CV with the ethnic last name of interest

Additionally, we will also analyse ethnic discrimination by comparing key outcomes for minorities and Anglo-Celtic CVs. We will compare a given minority CV to the rest of the CVs with an Anglo-Celtic name of the same gender **within** the same control group (assume for the example below that the minority name of interest is a female in control group 1).

Given n_c reviewers in control group 1, we will estimate the following general model for a given outcome of interest Y_i for reviewer i, and for all CVs of the same gender in control group 1 as the gender of the minority name of interest:

$$Y_i = \beta_0 + \beta_1 Minority_i + \varepsilon_i$$

$$\text{where } i = 1, 2, \dots, n_c$$

$Minority_i$ is an indicator for whether a particular CV has the name of the ethnic minority of interest, and ε_i is the error term of the model. Our key parameter of interest is β_1, as it captures the effect of having the minority name on our key outcomes of interest, which are:

- % CVs with female name in reference control group that are potentially suitable for job
- Average score (on a 5-10 cardinal scale) of potentially suitable CVs with female names in reference control group
- % CVs with female names in the reference control group that are shortlisted
- Average shortlist ranking (on a 1-5 ordinal scale) of CVs with female names in reference control group

Secondary Outcomes of interest

In addition to the models described above, we will also estimate models that include a set of control variables of interest. These variables include demographic characteristics of reviewers as well as a range of other variables that previous research has identified as being related to the bias displayed against females and/or minorities candidates.

To illustrate our approach to incorporating these control variables, we will specify the model for the pure gender discrimination effects described above; similar models will be specified for all other primary outcomes of interest related to the effect of de-identification and ethnicity. Using the notation described above, the new model is:

$$Y_i = \rho_0 + \rho_1 C_i + X_i \rho_2 + \omega_i$$

$$\text{where } i = 1, 2 \dots, n_{c1}, n_{c1} + 1, n_{c1} + 2, \dots\dots, n_{c1} + n_{c2}$$

Y_i is a key outcome of interest, C_i is an indicator for whether a reviewer i is in the reference control group, X_i is a vector containing control variables, and ω_i is the error term of the model. The key parameter of interest is ρ_1, and this is the average treatment effect conditional on the range of characteristics contained in X_i. We will estimate different models by gradually adding control variables to the X_i vector, while also potentially incorporating interaction terms between control variables. The controls we will potentially incorporate in the conditional models are:

- Age (categorical)

- Gender (binary)
- Identifies as LGBTI (binary)
- APS job level (categorical)
- APS agency (categorical)
- Ongoing disability (binary)
- Indigenous status (binary)
- Non-English speaking background (binary)
- APS Role (categorical—e.g. Corporate or non-Corporate role)
- How efficient the shortlisting process in the exercise is compared to the standard shortlisting process (categorical scale)
- Whether there are opportunities to improve APS recruitment practices (binary)
- The difference between a reviewer’s confidence that he/she shortlisted the best candidates and that other reviewers completing the exercise shortlisted the best candidates (categorical scale)
- Time since last meal (categorical scale)
- Intensity of current workload (categorical scale)

Appendix 4: Summary tables for reporting primary results

[Refer to pre-analysis plan published on the American Economic Association (AEA) Social Science Registry: https://www.socialscienceregistry.org/trials/1783]