

Development of a Bayesian Thurstonian Model for Analysing Ranking Data From Live Postdoc Recruitment

Noam Tal-Perry^{1,2,3}, Lara Abel^{1,2}, Mollie Etheridge^{1,2,4}, Jessica Hampton^{1,2}, Becky Ioppolo^{1,2}, Adrian Barnett⁵, Timothy R. Johnson⁶, and Steven Wooding^{1,2*}

¹Research Strategy Office, University of Cambridge, Cambridge, CB2 1TN, UK

²Bennett Institute for Public Policy, University of Cambridge, Cambridge, CB3 9DT, UK

³Murray Edwards College, University of Cambridge, Cambridge, CB3 0DF, UK

⁴Faculty of Education, University of Cambridge, Cambridge, CB2 8PQ, UK

⁵School of Public Health & Social Work, Queensland University of Technology, Brisbane, Australia

⁶Department of Mathematics and Statistical Science, University of Idaho, 875 Perimeter Dr, Moscow, ID 83844, USA

* Corresponding Author


Noam Tal-Perry  <https://orcid.org/0000-0003-2521-9546>

Lara Abel  <https://orcid.org/0000-0001-6359-0843>

Mollie Etheridge  <https://orcid.org/0000-0003-0589-9222>

Jessica Hampton  <https://orcid.org/0000-0001-6871-2846>

Becky Ioppolo  <https://orcid.org/0000-0003-1301-7947>

Timothy R. Johnson  <https://orcid.org/0000-0002-2600-5972>

Adrian Barnett  <https://orcid.org/0000-0001-6339-0374>

Steven Wooding  <https://orcid.org/0000-0002-8036-1054>

Correspondence concerning this article should be addressed to SW, email:
sw131@cam.ac.uk.

Abstract

Various academic stakeholders within and outside the UK have expressed their interest in the use of Narrative Curriculum Vitae (NCV) in funding and recruitment processes to address multiple research culture concerns, such as the improvement of equity, diversity, and inclusiveness in assessment, acknowledgement of a wider range of outputs and research activities, and promotion of non-linear career paths. A recent pilot study using a randomised clinical trial in live recruitment of postdoctoral candidates at the University of Cambridge examined the effect of CV format on panellists' ranking of applicants. For this, each applicant was asked to submit both a Standard CV (SCV) and a NCV when applying for the post, with CVs pseudo-randomly assigned across panel members before applications were ranked by each. Here, we use the data collected in the pilot study to construct a multi-layered generative model of the recruitment data, simulating applicants, panel members, and recruitments. Using the generative model, we created several simulation experiments, each varying one of the model parameters, and repeated the process a thousand times for each experiment. The resulting synthetic data was then modelled using a Bayesian Thurstonian model, a model apt for the analysis of ranked data coming from multiple raters, to test whether parameters were correctly recovered and how the change in parameters affected the dispersion of model estimates. The results from this process were used to inform the design of a subsequent experiment in a larger sample.

Keywords: Thurstonian model; Bayesian models; Generative model; Recruitment; Postdoc; Narrative CV

Introduction

General introduction

Over the last decade, policymakers and universities across the UK have identified a set of sector-wide issues concerning current research culture practices (Gottlieb et al., 2021; Russel Group, 2021; Universities UK, 2019; Wellcome, 2022; Wellcome Trust, 2020). The Concordat to Support the Career Development of Researchers (2019) highlights the need for a diverse workforce and inclusive research culture at UK Universities

To foster equality, diversity, and inclusion (EDI) within the academic system while acknowledging a wider range of research outputs and activities without disadvantaging non-linear career paths, the UK Research and Innovation (UKRI, the primary government funder of research in the UK) published their People and Teams Action Plan for institutions to achieve these goals (UKRI, 2023). Among the proposed actions are the introduction of Narrative Curriculum Vitae (NCV) in funding applications, and their potential use for academic job recruitment and promotion processes.

The NCV

In its standard version, the curriculum vitae (SCV) lists an applicant's education and research experience in a bullet point format, highlighting the institutions from which degrees were awarded, their academic positions, and the track record of authored publications. This mode of presentation was criticised as putting too much emphasis on narrow achievements of papers and grants, often assessed through academic journal publications, thus narrowing the range of skills that are incentivised and excluding many researchers who make valuable contributions to the scientific community (Curry et al., 2022; Strinzel et al., 2021). Additionally, the SCV encourages readers to focus on a limited range of outputs, typically peer-reviewed publications published in high-ranking journals, although these may not predict the quality of work (Pietilä et al., 2023; see also the San Francisco Declaration on Research Assessment (DORA), <https://sfdora.org/read/>). It was additionally suggested that the SCV format leaves little or no space for researchers coming from underrepresented groups to demonstrate their participation in academic activities that promote equity, diversity, and inclusion (EDI) in science which are often engaged by underrepresented researchers, thus limiting the acknowledgement given to these contributions and preserving the lack of representation of these groups (Bhalla, 2019).

Recently, the NCV has been promoted 'to capture the outputs of those not on a standard academic career pathway' (UKRI, 2023, p. 32). There are currently several NCV formats that are being experimented with by various countries and research organisations, with a common purpose to highlight and contextualise these broader contributions which may not so easily be captured and quantified in the SCV. The Résumé for Research and Innovation (R4RI, see <https://www.ukri.org/apply-for->

[funding/how-to-apply/resume-for-research-and-innovation-r4ri-guidance/](#)) is a NCV template introduced by UKRI. The R4RI format is structured around four modules, each focusing on a specific type of contribution the applicant made to the scientific community, while its narrative form allows to contextualize those contributions and explain their importance (Adams et al., 2023; Strinzel et al., 2021). Other major research and funding agencies across the UK, such as Alzheimer's Research UK and Cancer UK, have already adopted the R4RI format for their funding calls (UKRI, 2021), and research agencies in other countries are experimenting with other NCV format versions, including the Dutch Research Council, Luxemburg National Research Fund, and the Swiss National Science Foundation (Fritch et al., 2021).

Research on NCV in recruitments

Thus far, the NCV has been primarily used and investigated in the context of funding applications, with little work done on the use of NCV in recruitment (for a review, see Bordignon et al., 2023). One key difference between those uses is the role of covering letters, which play a major part in job applications and may overlap in their content with NCVs (Ioppolo et al., 2024). Previous work on the feasibility and usefulness of using the NCV format in recruitment has often studied practitioners' views on the NCV format, the availability of NCV resources, and suggestions for how the NCV format could be improved. For example, Aubert Bonn et al. (2024) used a workshop format to elicit organisations' and researchers' views on the NCV; Meadmore et al. (2022) followed a mixed methods approach including secondary analysis of existing NCV documents and guidelines combined with interviews and focus groups with relevant stakeholders. To our knowledge, only one study tested the usefulness of the NCV format in the academic recruitment of early career researchers (Adams, 2021).

In a previous pilot study, we explored whether the NCV format changes who is shortlisted, and if so, in what way it changes, by setting up a randomised controlled trial (RCT) to test the effect of CV format (SCV vs NCV) on recruitment outcomes. We tested the design of our study in a sample of five live postdoc recruitments at the University of Cambridge, while collecting qualitative data through interviews with the applicants and the hiring panel, as well as quantitative data in the form of ranking of the applications. The qualitative results observed in this pilot study were discussed in a separate publication (Ioppolo et al., 2024).

Generative model

Here, we use the quantitative data obtained from the pilot study to inform a generative model simulating different scenarios we may obtain from the main phase experiment. A generative model uses distributions of existing data to generate new synthetic data, which is similar to the existing data. The synthetic ranking data produced by the generative model was then modelled using a Bayesian Thurstonian model. By systematically changing different parameters of the design, we tested whether the Thurstonian model could recover the original parameters used to generate the synthetic data, and additionally to test our assumptions regarding the data and to study the effect of varying aspects of our experimental design on the expected results.

The goal of the current paper is to describe the generative model used for our study and to explore through simulations how various key parameters of this design, such as the number of live recruitments, panel members, and applicants, affect the model's estimates. The results of these simulations inform design choices for our main phase study and test the feasibility of our analysis approach given the design choices made.

Method

Participants

The data informing our generative model were generated in the pilot phase of our experiment on the use of NCVs in academic recruitment (Ioppolo et al., 2024). In this study, principal investigators (PIs) recruiting for postdoctoral positions at the University of Cambridge were approached to take part in the study. A total of five PIs signed up for participation, resulting in a total of five live recruitments for postdoctoral positions taking place between March and August 2023, all from the Science, Technology, Engineering, and Maths (STEM) disciplines. Each recruitment panel consisted of 2-5 panel members including the PI (mean 3 ± 1 standard deviations [SD] panel members per recruitment), totalling 17 participating panel members, of which 15 complied with the experimental procedure and were included in the analysis (see **Table 1**). All panel members gave written informed consent to participate. Panel members did not receive compensation for their participation. A total of 148 applicants (mean 30 ± 16 SD applicants per recruitment, range 14-56 applicants per recruitment) submitted an application to one of the positions. All applicants were invited via email to participate in

	Subject	Panel members	Total applicants	Consenting applicants	NCV submissions	Consenting NCV submissions
01	Allied health	4	22	9 (40%)	7 (32%)	6 (67%)
02	Material and technology	3	56	26 (46%)	15 (27%)	10 (38%)
03	Medical sciences	2*	31	15 (48%)	4 (13%)	3 (20%)
04	Physics and astronomy	3	25	9 (36%)	8 (32%)	5 (56%)
05	Mathematical sciences	5†	14	5 (36%)	2 (14%)	1 (20%)
Total (%)		17	148	64 (43%)	36 (24%)	25 (39%)

Table 1 *Recruitment statistics.* The table depicts the subject and number of panel members, applicants, consenting applicants, and compliance with narrative CV (NCV) guidelines for each recruitment in the study. Subjects were recoded according to the second level of the UK Higher Education Statistics Agency (HESA) Common Aggregation Hierarchy. Reported percentages for consenting applicants and NCV submissions are calculated in proportion to the total number of applicants in the recruitment, while consenting NCV submission percentages are calculated in proportion to the number of consenting applicants. *One panel member did not comply with instructions regarding the 2nd ranking and was excluded from that analysis. †One panel member did not comply with instructions and was excluded from all reported analyses.

the study (participant information sheet available in online materials) once the application had closed. By consenting to the study, participants gave permission to the research team to analyse their personal data and all information pertaining to their application. All applicants were told that participation in the study would not affect their chances of being considered for the position and that panel members would not be informed about their decision. Consenting participants were offered a feedback session on their NCV from a careers consultant working with the project as compensation for participation. A total of 64 applicants (mean 13 ± 8 SD applicants per recruitment, range 5-26 applicants per recruitment, making a total of mean $41\% \pm 6\%$ SD of total applications per recruitment, range 36-48% of applicants per recruitment) gave informed consent to participate in the study and were included in the analysis. This study was carried out in accordance with the guidelines and regulations set by the Declaration of Helsinki and was approved by the Research Ethics Committee of the Department of Psychology at the University of Cambridge (reference number PRE.2022.071).

Procedure

The job adverts listed on the University of Cambridge website for participating recruitments included additional instructions and Further Particulars pertaining to the study. Applicants were instructed to submit a NCV in addition to their SCV and all other requested materials (such as a cover letter) when submitting their application for consideration, regardless of participation in the study. This was to ensure that applicants not wishing to participate in the study would not be negatively impacted by the lack of submission. To assure all applicants have sufficient knowledge of the new requested format, the Further Particulars included a reference to online materials instructing how to prepare a NCV. Applicants were informed that the decision regarding the application will be based on both types of CVs. A total of 36 applicants complied with the instructions and submitted both types of CVs (mean 7 ± 5 SD applicants per recruitment, range 2-15 applicants, making a total of mean $24\% \pm 9\%$ SD of applicants per recruitment, range 13-32% of applicants per recruitment). Of these, 24 applicants also consented to participate in the study (mean 5 ± 4 SD of applicants per recruitment, range 1-10 applicants per recruitment, making mean $65\% \pm 16\%$ SD of consenting applicants per recruitment, range 50-86% of consenting applicants per recruitment). The total number of NCV submissions per recruitment is depicted in Table 1.

Following the application closing date, the CV type (standard/narrative) for each applicant that submitted both types of CVs was pseudo-shuffled between panel members, such that each panel member received only one type of CV at the initial stage, while ensuring half of the panel members (rounding down) receive each type of CV for a given applicant. Where applicants did not submit a NCV, the SCV was given to all panel members. Panel members also received all additional materials the applicant included in their application (such as a cover letter). Each panel member was then asked to individually rank all credible candidates according to the materials they received, aiming for 10-12 applicants, in reverse order (such that 1 indicates the best applicant) and without using ties. We will refer to this ranking in this paper as the 1st ranking set. The use of ranking instead of scoring has the benefit of simplifying the assessment process and is more analogous to the explicit or implicit process that takes place in normal recruitment. The ranking instructions were designed such that more applicants than

would reasonably be shortlisted were ranked to create more data for accurate estimation, but without requiring to distinguish between applicants who were not credible, with the restriction of untied ranks stemming from our modelling approach (see Statistical Modelling section). In total, 67 of the applicants (regardless of consent) were ranked by at least one panel member in the first ranking round (mean 13 ± 4 SD applicants per recruitment, range 7-18 applicants per recruitment), suggesting that $53\% \pm 29\%$ SD of applicants per recruitment (range 32%-100% of applicants per recruitment) were considered credible by at least one panel member. The number of ranked applicants per panel member in the first ranking round ranged from 6-14 (mean min rank 9 ± 3 SD per recruitment, mean max rank 11 ± 3 SD per recruitment). Once panel members submitted their ranking to the research team, they were given the full applications made by each applicant, thus receiving the second CV type for applicants who submitted both types of CVs. In cases where applicants did not supply a NCV, panel members received the same application material in this second round. Panel members were then asked to individually provide a 2nd ranking set based on the new material. The analysis of the 2nd ranking set is outside the scope of the current study, which will focus on the 1st ranking set. Only after submitting this second set, panel members were able to confer with each other, at which point the recruitment proceeded as normal, with applicants being shortlisted, invited for interview, and made an offer.

Generative model

The data from the pilot study were used to inform a generative model of rankings in the first ranking round. At the core of this generative model was a Thurstonian model (Thurstone, 1927, 1931), which assumes that the observed ranking frequency is consistent with the probability distribution of values on a latent (unobserved) continuous scale (see **Figure 1** for an illustration of the model). In this case, the latent score represents the panel member's perception of the applicant's true suitability for the role, such that applicants with a high latent score are more suitable than those with a lower latent score. Ranking is then produced by each panel member according to the ordered sampled latent scores across applications. Inferences for the parameters of the underlying distribution of scores can then be made based on the rankings. By introducing additional covariates to the model (e.g., observed CV type), we can estimate their influence on the latent scores.

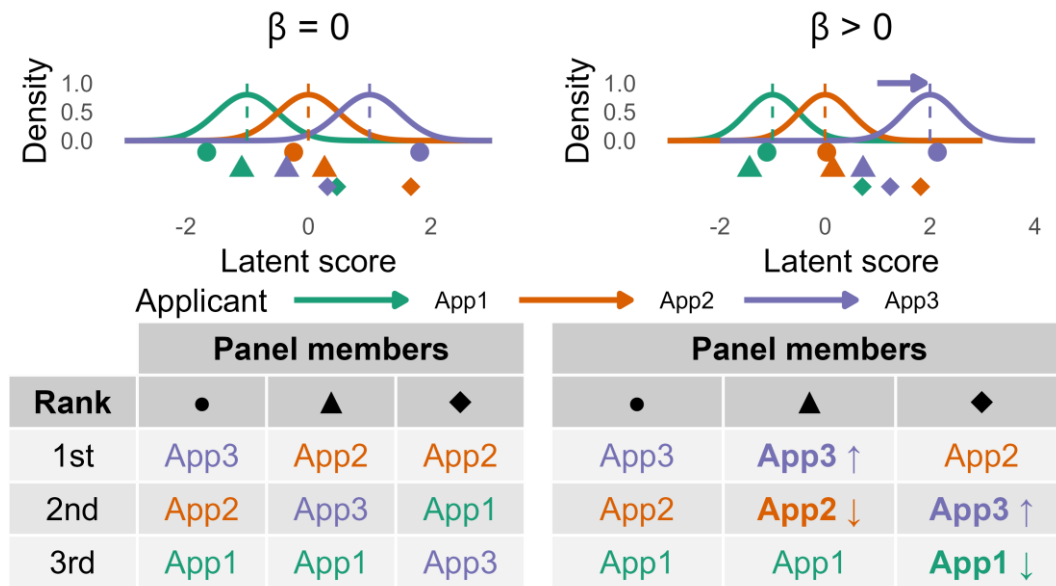


Figure 1 *Thurstonian model illustration.* The plots in the upper panels depict the latent score distributions of three applicants (coloured) differing in their latent skill (dashed lines). Three panel members (shapes) assess each applicant by sampling from the corresponding distribution (indicated by the coloured shapes). The sampled latent scores are then reversed-ranked by each panel member (bottom tables). The plot on the left depicts a scenario where the studied covariate (e.g., NCV) does not affect the latent score ($\beta = 0$). The plot on the right shows the same data, but an effect for the studied covariate ($\beta > 0$) on the latent score of Applicant 3 (shift designated by arrow), leading to higher sampled latent scores for each panel member, which translates to a change in the resulting ranks for two of the panel members (bottom right table).

Our generative model (see **Figure 2**) consists of a simulation of recruitments (default number: 40), within each we generated the number of panel members (truncated normal distribution, mean 3.2, SD 0.7, boundaries 2-5) and the number of applicants (truncated normal distribution, mean 30, SD 16, boundaries 10-75) based on the frequencies observed in the pilot study. For each recruitment, we generated a recruitment-level binary covariate (e.g., representing whether this recruitment was for a STEM or Arts, Humanities, and Social Sciences position), and for each panel member, we generated a panel member-level binary covariate (e.g., representing whether that panel member had previous experience with NCVs), with covariates generated using a Bernoulli distribution with a probability of .5. Bernoulli covariates were used for simplicity, and their parameters values were not informed, but were subsequently tested through simulation (see Simulations section). For each applicant, we generated a latent score assuming a normal distribution an arbitrary mean of 0 and SD of 1. We then assigned whether applicants submitted a NCV based on the ratio observed in the pilot

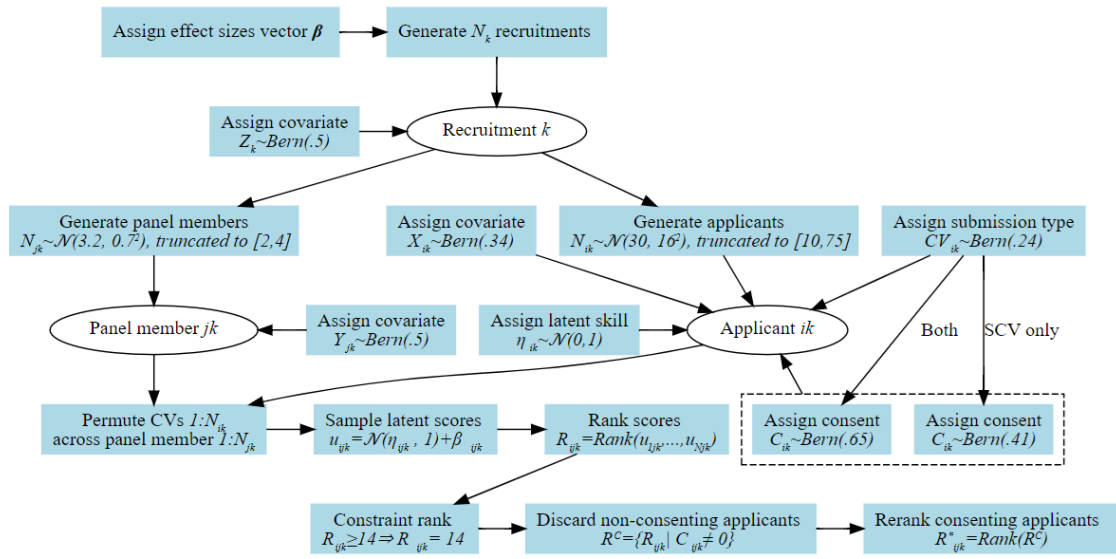


Figure 2 *Generative model.* The diagram illustrates the functions (blue boxes) and objects (white ovals) used in the generation of our data, with values informed by the pilot study.

study (Bernoulli distribution with a probability of .24) and whether they consented to participate in the study using different consent odds according to submission type, as observed in the pilot study (Bernoulli distribution with probability of .41 and .65 for SCV and both submissions, respectively). We additionally generated for each applicant a binary applicant-level covariate (e.g., applicant gender), based on the gender ratio we found in our pilot study (Bernoulli distribution with a probability of .34).

Next, replicating our experimental procedure, we used constrained permutation of CV allocation to balance the CV type for each simulated panel member for applicants submitting both types of CVs, with the SCV being allocated to the remaining applicants. Each simulated panel member then sampled from the applicant’s latent score distribution corresponding to the assigned CV type, with distribution assumed to be normal with a mean centred on the applicant’s latent score and a SD of 1 arbitrary unit. For simplicity, we assumed the SD of the distribution to be identical across applicants, panel members, and between CV types. Applicants were then ranked based on the sampled latent scores for each panel member. To simulate the instruction to rank only credible applicants, we constrained the number of ranked applicants in each simulated recruitment by producing a max rank set as a fraction of the number of applicants, based on the values observed in the pilot data (53% of applicants per recruitment). We

additionally set a ceiling based on the max ranked applicant in our pilot data (14), simulating our instructions to aim for 10-12 ranked applicants. All applicants beyond the lowest-ranked applicant were assumed to tie for the lowest rank. We then removed the non-consenting applicants from the max rank set and re-ranked the remaining applicants by their original ranks (e.g., if an applicant ranked 2 was removed from a set of four applicants ranked 1, 2, 3, and 3, the applicants originally tied for lowest rank 3 were now re-ranked as tied for 2), to simulate the data available for the study.

Statistical modelling

Inferences from the ranking data from the generative model were made using a Bayesian Thurstonian model (Johnson & Kuhn, 2013; Yao & Böckenholt, 1999), which assumes that the observed ranking can be described as the probability of ranking of an unobserved continuous latent skill (see Figure 2). In our study, the model assumes that each applicant's latent score can be described as:

$$\eta_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1 x_3 + \beta_5 x_1 x_4 + v_i \quad (1)$$

Where:

- η_i stands for the latent skill of applicant i
- β_1 stands for the main effect for CV type x_1 (standard/narrative)
- β_2 stands for the main effect for an applicant-level covariate x_2 (e.g., their gender)
- β_3 stands for the interaction term between CV type (x_1) and the applicant-level covariate (x_2)
- β_4 stands for the interaction term between CV type (x_1) and the panel member-level covariate x_3 (e.g., seniority)
- β_5 stands for the interaction term between CV type (x_1) and the recruitment-level covariate x_4 (e.g., discipline)
- v_i stands for the applicant-specific random effect of applicant i

And with the β coefficients assumed to be normally distributed with a weakly informed prior at ($\mu = 0, sd = 2$) and v assumed to be normally distributed with $\mu = 0$ and $SD = 1/\tau$, with precision parameter τ (the inverse of variance) being gamma distributed with a prior on shape parameter $\kappa = 2$ and scale parameter $\eta = 2$. The choice of prior was

based on the distribution of the latent skill in the simulated population ($N(0,1)$ arbitrary units), with the assumption that effects sizes are expected to be modest.

Of interest to the research question, the β coefficient could address whether CV type affects ranking (β_1), and whether certain CV types benefit or disadvantage participants coming from certain backgrounds (β_3 interaction term) or disciplines (β_5 interaction term), as well as whether the rating panel member's background affects results (β_4 interaction term). This simplified model can be extended to any number of applicant-, panel member-, or recruitment-level covariates. Given the experimental design, the main effects for panel member- and recruitment-level covariates have a no influence on rankings, rendering them unidentifiable and thus they were excluded from this model.

The model assumes that the perception of each panel member j assessing applicant i is normally-distributed and centred on the applicant's latent score with an SD of 1 (arbitrary unit), resulting in the applicant's perceived latent score by that panel member (utility) u :

$$u_{ij} = N(\eta_i, 1) \quad (2)$$

The current implementation of the model does not permit tied ranking beyond the last ranking.

Unlike common applications of the Thurstonian model (see Johnson & Kuhn, 2013), in our study the stimuli to be ranked (i.e., applicants) can be different for each ranking (i.e., recruitment). Thus, correlations between the scores of the applicants are not estimable, simplifying the model significantly, but their variability can be estimated.

The modelling was performed using the R programming language and the JAGS statistical language (see Software section). Modelling was done using four Markov Chain Monte Carlo (MCMC) chains taking 10,000 samples from the posterior distribution for each of the model coefficients after discarding the first 1000 samples (burn-in phase). Diagnostic checks were performed on the base model using parameter values informed by the pilot study.

Simulations

To explore the behaviour of the generative model, we ran ten simulation experiments, each systematically manipulating one of the design parameters, while setting the rest of

the parameters to their default values as informed by the pilot data. Each simulation experiment was explored through 1000 simulated studies per parameter level, with each study including a varying number of recruitments. The results from each simulated study were modelled using the Bayesian Thurstonian model described above. For each Bayesian simulation, we extracted the point estimates (mean, median, maximum a-posteriori probability [MAP]) and dispersion estimates (SD, median absolute deviation [MAD], mean squared error [MSE], and 50%, 75%, 89% and 95% highest density region [HDR]) of the posterior distribution for each of the model's five β coefficients and the σ parameter and averaged these across simulations. A seed was set for each simulation for reproducibility.

The following simulation experiments were conducted: a) varying the number of recruitments between 20 and 60 in increments of 10; b) varying the number of panel members per recruitment between 2 and 6; c) varying the number of applicants per recruitment between 25 and 65 in increments of 10; d) varying the rate of NCV submissions between .1 and .5 in increments of .1; e) varying the consent rate between .1 and .5 in increments of .1, regardless of submission type; f) varying the percent credible (ranked) applicants out of the recruitment between .1 and .5 in increments of .1; g) varying the max number of ranked applicants between 3 and 15 in increments of 3; h-j) varying the binomial probability ratio of the recruitment-, panel member, and applicant-level covariates between .1 and .5 in increments of .1.

Software

The generative model and the analyses of the results were scripted using R v4.4.1 (R Core Team, 2023) on RStudio v2024.04.0-735 (Posit team, 2024). The Bayesian Thurstonian model was modelled using JAGS v4.3.2 (Plummer, 2003) on R using the `rjags` package (Plummer, 2023) with parallel computing done via the `dclone` package (Sólymos, 2010). Extraction of posterior point and dispersion estimates was done using the `bayestestR` package (Makowski et al., 2019).

Data availability statement

The script for the generative model, the results from the simulations, and the script producing the graphs are all freely available online on the project's Open Science Foundation (OSF) repository at <https://doi.org/10.17605/OSF.IO/DKW29>. The

Participant Information Sheet and the job advertisement template can be found in our previous publication OSF repository at <https://doi.org/10.17605/OSF.IO/GWA9R>.

Results

Effect size recovery and simulation

We began by testing whether the model can correctly recover the parameters used in the generative process. We set the generative model's parameters to their default values as determined by the pilot data (see Methods section for the list of values) and varied the effect size for each of the five β coefficients independently. Additionally, we assessed the model's ability to recover the σ parameter (inverse of precision) by varying the distribution width at the population-level (simulating the underlying suitability distribution of applicants) and at the panel member observation level (simulating different precision of ability to accurately assess applicants), and tested how these affect the β coefficients dispersion estimates.

Means and MSEs obtained from the simulations showed that the model correctly recovered all five β coefficients across a range of tested values (see **Figure 3**), with a slight increase in error with increasing effect sizes for the CV type (β_1) and applicant covariate (β_2) main effects. Similarly, the model successfully recovered the σ coefficient when varying the population-level SD used in the generation of applicants' base levels (see **Figure 4**, left panel), with increasing error for increasing simulated values. As expected, increasing the population-level SD increased the uncertainty surrounding the β estimates (see **Figure 4**, right panel), affecting linearly the applicant covariate and non-linearly the CV type covariate and the interactions with it. Manipulation of the panel member observation-level SD (**Figure 5**, left panel) recovered the inverse of σ (i.e., the τ or precision parameter), with increasing underestimation with increasing simulated values. Decreasing precision led to a non-linear increase in the uncertainty levels surrounding β estimates, growing more quickly for the CV type coefficients (**Figure 5**, right panel). Results for the other extracted point estimates (median and MAP) followed a similar distribution and are reported in **Supplementary Figure S1-S2** for the β and σ coefficients, respectively.

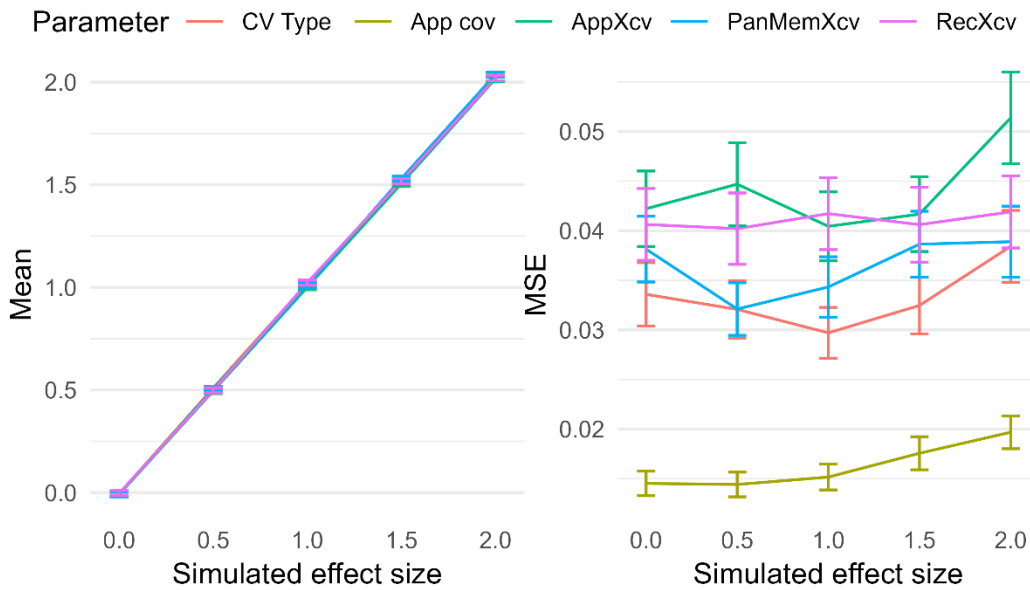


Figure 3 *Effect size recovery.* Mean (left) and mean squared error (MSE, right) for each of the model’s five β coefficients (coloured lines), for effect sizes set between 0-2 (x-axis), averaged across 1000 simulations per parameter and effect size, with each simulation consisting of 40 simulated recruitments. Error bars depicting 95% confidence interval around the mean. CV Type = CV type main effect (β_1); App cov = applicant covariate main effect (β_2); AppXcv = applicant covariate X CV type interaction effect (β_3); PanMemXcv = panel member covariate X CV type interaction effect (β_4); RecXcv = recruitment covariate X CV type interaction effect (β_5).

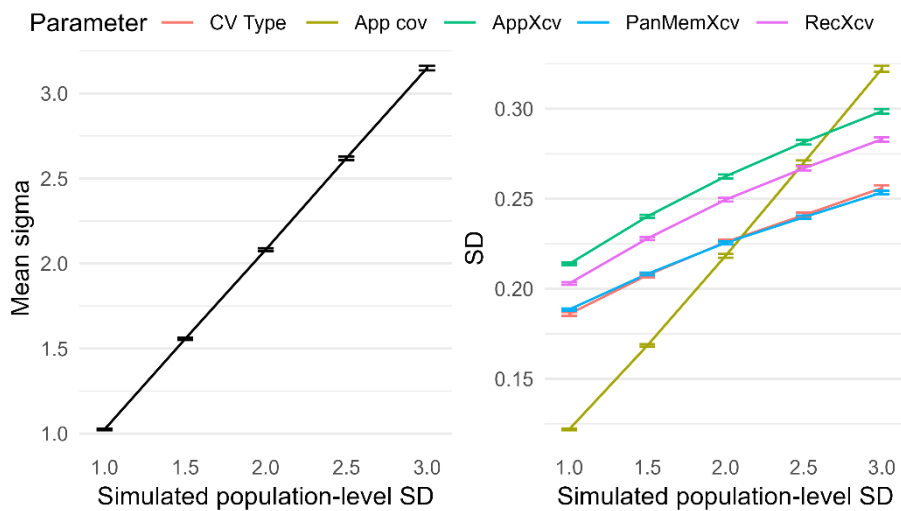


Figure 4 *Effect of Population-level standard deviation (SD).* Mean σ parameter recovery (left) and SD of each of the model’s five β coefficients (in coloured lines, right) for simulated population-level distribution widths set between 1-3 SDs (x-axis), averaged across 1000 simulations per parameter and level, with each simulation consisting of 40 simulated recruitments. Error bars depicting 95% confidence interval around the mean.

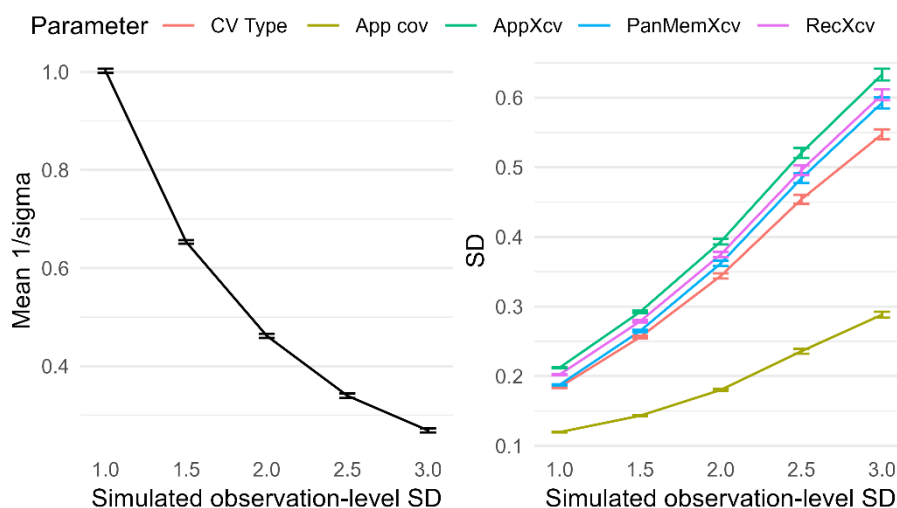


Figure 5 Effect of panel member observation-level standard deviation (SD). Mean recovered precision parameter ($\tau = 1/\sigma$, left) and SD of each of the model's five β coefficients (in coloured lines, right) for simulated observation-level distribution widths set between 1-3 SDs (x-axis), averaged across 1000 simulations per parameter and level, with each simulation consisting of 40 simulated recruitments. Error bars depicting 95% confidence interval around the mean.

Number of recruitments, panel members, and applicants

We aim to study a total of 40 recruitments in the main phase of our study, requesting PIs to enlist at least two additional panel members to rank the applications submitted for the advertised position. According to the pilot data and from our examination of HR data from the University of Cambridge, the number of applications submitted to research positions varies greatly between recruitments, with an average of 30 applications per recruitment in our pilot data, regardless of consent and instructions adherence rates. To test how these three factors affect the degree of uncertainty surrounding parameter estimation, we conducted three separate simulation experiments: a) fixing the number of recruitments while letting the number of panel members and applicants per recruitment vary; b) fixing fixed the number of panel members per recruitment while setting the number of recruitments at 40 and allowing the number of applicants to vary; and c) fixing number of applicants per recruitment while setting the number of recruitments to 40 and letting the number of panel members vary. For each experiment, we extracted the posterior distribution dispersion estimates for each of the model's coefficients to assess the effect of the manipulation on their dispersion.

As expected, increasing the sample size for each of those design parameters decreased the overall dispersion, as observed in the posterior SD results (see **Figure 6**; similar results obtained for MAD, see **Supplementary Figure S3**), with decreasing gains with linearly increasing parameter values. Dispersion estimates for CV type (β_1) and its

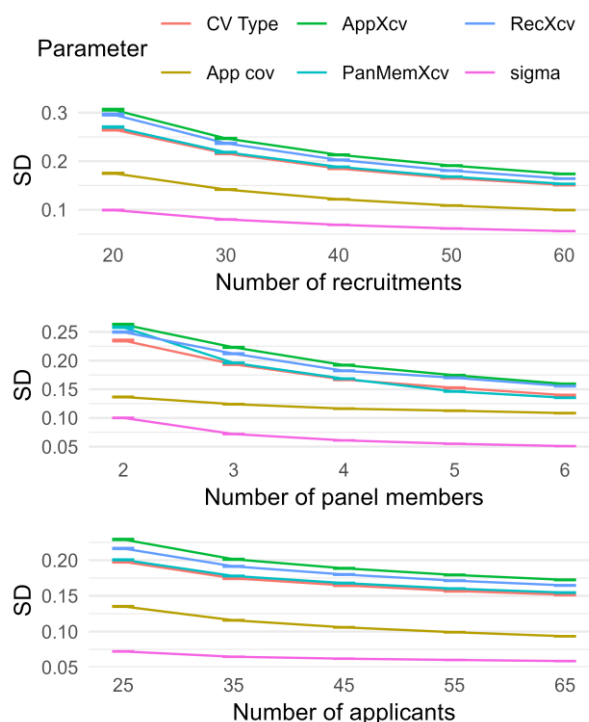


Figure 6 Effect of sample size manipulations on the dispersions of the model's coefficients. The mean posterior standard deviation (SD) for each of the model's coefficients (coloured lines), averaged across 1000 simulations, as a factor of varying the number of recruitments in the simulation (upper panel), the number of panel members per recruitment (middle panel), and the number of applicants per recruitment (lower panel). Error bars depicting 95% confidence interval around the mean.

corresponding interactions with the other covariates ($\beta_3 - \beta_5$) were larger than the dispersion estimates surrounding the applicant covariate effect coefficient (β_2) and σ coefficient. Changes in the number of panel members (Figure 6, middle panel) had little effect on the applicant covariate coefficient, while changing the panel size from 2 to 3 members had a relatively large effect on the interaction between panel member covariate and CV type coefficient (β_4). Overall, increasing the number of recruitments had the most profound effect on decreasing dispersion estimates, while increasing the number of applicants per recruitment had the least.

NCV submission rate, consent rate, credibility rate, and max ranked applicant

In our pilot study, the number of applicants that ended up in the analysis was lower than the number of applicants that applied for the position. First, only 41% of applicants per recruitment agreed to participate in the study and were included in the analysis, with those submitting NCV tending to consent at higher rates, standing at 65% of consenting applicants per recruitment. Second, we asked panel members to rank as many credible applicants as they could, which resulted in 53% of applicants per recruitment

(regardless of consent) being ranked by at least one panel member. Third, we asked panel members to rank beyond the number of applicants they would shortlist and to aim to rank about 10-12 applicants, with pilot data showing a maximum of 14 applicants ranked in a single recruitment by an individual panel member. Lastly, the CVs the panel members ranked were not uniformly distributed among the two types, as only 24% of applicants per recruitment (regardless of consent and being ranked) adhered to the instructions and submitted both types of CVs. Each of those factors affected the number of ranked observations for each CV type. To examine how these factors affect the uncertainty surrounding the estimation of the parameters, we examined the posterior dispersion estimates under four separate simulation experiments, varying a) the NCV submission rate per recruitment; b) the consent rate per recruitment, with consent being constant across submission types; c) the credible applicants rate amongst all applicants in each recruitment; and d) the maximum number of ranked applicants per panel member, simulating changes in ranking instructions.

Extracted posterior SDs are depicted in **Figure 7** (see **Supplementary Figure S4** for similar MAD results). As with the sample size manipulation (see Figure 6), the dispersion estimates were generally larger for coefficients related to CV type ($\beta_1, \beta_3 - \beta_5$) than the coefficient for applicant covariate main effect (β_2) and the σ coefficient. Across all these four design parameters, increasing the consent rate (Figure 7, upper right panel) had the largest effect on decreasing uncertainty surrounding the coefficients.

As expected, varying the rate of NCV submission (Figure 7, upper left panel) did not affect the applicant covariate main effect. We additionally observed a dramatic decrease in dispersion estimates when moving from the NCV submission rate of 0.1 to 0.3, which was similarly observed for the consent rate (upper right panel) and the credibility rate (lower left panel). This could potentially be explained by the effect of increasing the number of valid recruitments, that is, recruitments where there were at least two ranked and consenting participants that would allow the variance to be estimated.

Surprisingly, the effect of credibility rate on dispersion remained relatively flat between the 30-90% range. This could partly be explained by the hard threshold set at 14 ranked applicants as part of the default design parameters. Thus, a recruitment that had at least 50 applicants would not benefit from the increase in the percent of credible

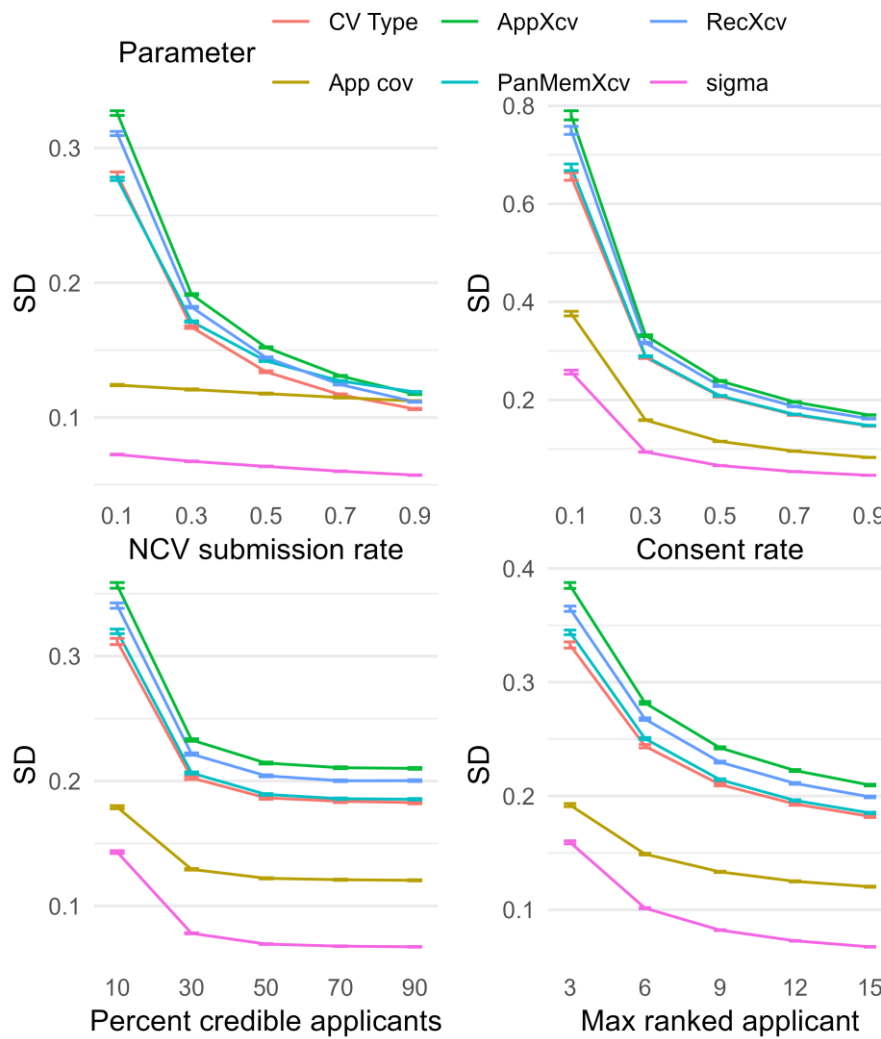


Figure 7 Effect of rates and rank manipulations on the dispersions of the model's coefficients. The mean posterior standard deviation (SD) for each of the model's coefficients (coloured lines), averaged across 1000 simulations and 40 simulated recruitments per simulation, as a factor of varying the rate of NCV submissions per recruitment (upper left), the consent rate per recruitment (upper right), the percent credible (ranked) applicants per panel member (lower left), and the maximum number of applicants ranked per panel member (lower right). Error bars depicting 95% confidence interval around the mean.

applicants past 30% ($50 \cdot 0.3 = 15$ ranked applicants). As the number of applicants varied between recruitments, the manipulation still affected a portion of the generated recruitments that had fewer applicants, with decreasing gains with increased percentages. Separately manipulating the maximum number of ranked applicants (lower right panel) showed that asking panel members to rank more applicants leads to decreasing dispersion surrounding the coefficient estimation across the range inspected, which would generally still fall below the set default number of ranked applicants per recruitment and panel member (53%).

Sample characteristics

Our experiment includes three types of covariates that describe the sample characteristics: an applicant-level covariate (e.g., their gender, ethnicities, geographic location, etc.); a panel member-level covariate (e.g., their seniority; gender, whether they are the recruiting PI, etc.); and a recruitment-level covariate (e.g., the discipline or faculty, the size of the department), each assumed to affect the applicant latent skill or the sampled latent skill while interacting with CV type. For simplicity, we assumed that the distribution of each of those covariates is binary and set the applicant-level covariate ratio Bernoulli probability at .34 according to the gender ratio we observed in the pilot study while setting the panel member- and recruitment-level covariates ratios Bernoulli probabilities at .5. To test the feasibility of the study to estimate effects of variables with varying degree of skewness, we conducted three simulation experiments, each varying one of the covariate ratios and measuring their effect on the posterior dispersion estimates.

As expected, the covariate ratio only affected the model's coefficient related to the covariate in question (see **Figure 8**), with similar results observed for MAD (see Supplementary Figure S5). We observed a non-linear relationship between the increase in the covariate ratio and the respective model coefficient in all three cases. Of interest, our results show that covariate ratios of .1 lead to particularly large uncertainty surrounding the coefficient estimation, and that the difference in posterior distribution width was negligible between covariate ratios of .3 and .5.

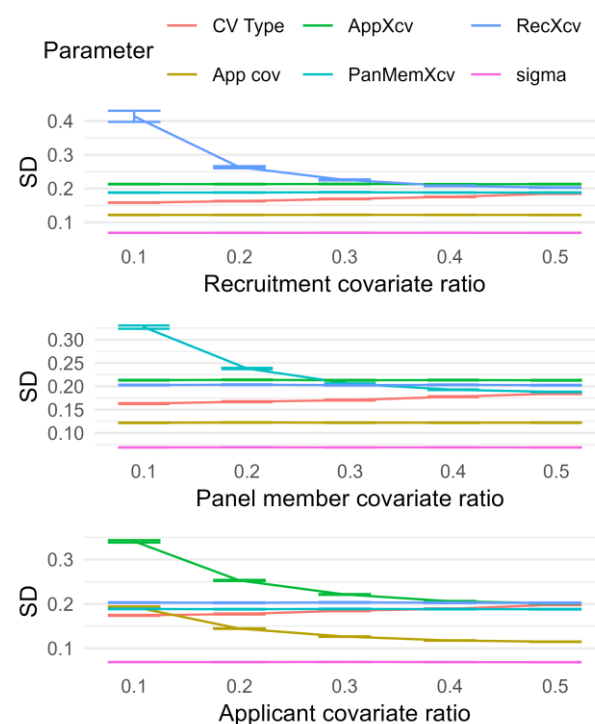


Figure 8 Effect of covariate ratios manipulations on the dispersions of the model's coefficients. The mean posterior standard deviation (SD) for each of the model's coefficients (coloured lines), averaged across 1000 simulations and 40 simulated recruitments per simulation, as a factor of varying the Bernoulli probability of the recruitment (upper panel), the panel member (middle panel), and the applicant (lower panel) covariate ratios. Error bars depicting 95% confidence interval around the mean.

Discussion

The purpose of the current study was to construct a generative model for a randomised clinical study on the effect of CV type on ranking in live recruitment for postdoctoral research positions at a higher research institution. We based our generative model on a pilot study conducted on 5 live recruitments at the University of Cambridge, whose results informed the parameter selection for the current model. The model was then used for a series of simulation experiments to test the effect of various design parameters and choices on the model estimates, with results being modelled using a Bayesian Thurstonian model.

In the following, we first summarise the main findings from this study and the conclusions we can draw for the main study on a larger sample. We then turn to discuss the model's assumptions and their potential effect on the observed results. Lastly, we discuss how this model can be extended and generalised to other cases.

Conclusions from the simulation experiments

The observed findings from our simulation experiments can inform several decisions on the design and conduct of the study's main phase.

Sample size. Results from the simulation of the sample size (see Figure 6) suggest that the number of recruitments had the largest effect on the reduction of uncertainty surrounding coefficient estimates. This can be expected, as even with two panel members and a relatively small number of applicants, each recruitment contributes numerous new observations for the model to base estimation on. The overlap in the parameter space between simulation experiments additionally allows us to make some direct comparisons. We can observe that 40 recruitments with 2 panel members each produce roughly similar dispersion estimates as 30 recruitments with a mean of 3.2 panel members. Similarly, it would take 40 recruitments with 5 panel members each to achieve dispersion estimates at the range that 60 recruitments would produce. In our pilot study, we observed that PIs, and especially junior PIs, could face difficulty in recruiting panel members. Thus, these results suggest the main study should focus on increasing the number of recruitments to achieve conclusive results.

An interesting observation is the effect of the number of panel members on the panel member-related covariate, which decreased substantially from increasing the number of panels from 2 to 3 panel members. This suggests the main study should

encourage PIs to aim for at least 2 additional panel members where possible, albeit recruitment with 2 panel members will have a contribution and should not be discarded. Lastly, the effect of the number of applicants was relatively negligible compared to the other two factors. Whilst this factor is outside the control of the researchers, it suggests that the main study should not concern itself with the number of applicants that applied for a position or particularly aim for disciplines that attract larger numbers of applicants.

Submission, consent, and ranking. The results on these factors (see Figure 7) indicated that the most substantial reduction in uncertainty could be achieved through the increase of consent rate. While this factor is eventually up to the applicants to decide, this suggests the study would benefit from attempts at convincing applicants to participate, for example by explaining its value, by making the consent information easy to read and comprehend, and/or by offering incentives for participation. As expected, the effect of adherence to the NCV submission guidelines mainly affected the model coefficients that are related to CV type, i.e. its main effect and its interactions with the covariates. The results show that this effect is highly non-linear, though the range of 0.3-0.9 was relatively linear. Given the low adherence rate observed in the pilot study (average of 24% per recruitment), the results suggest the main study would benefit from actions that would increase this rate, for example by simplifying the instructions or by offering additional incentives contingent on the submission of the NCV.

Findings from the number of credible applicants simulation experiment showed little difference in uncertainty levels once 30% of the candidates were considered credible. This is factored by the existence of the max ranking parameter (set at 14 for these simulations, based on the pilot studies), which limits the newly added information more credible candidates produce as they fall above this threshold and are considered to be tied for last rank. While this is another factor that falls beyond the control of the researchers, it does suggest that variation in perceived credible candidates between recruitments and panel members should largely have little effect on results. The findings from the simulation experiment on the maxed ranked applicant, which corresponds to the instructions given to the panel members, show an increasing benefit of ranked applicants across the inspected range. This indicates that applicants at the bottom of the ranking list are still beneficial for model estimation. While this suggests results could be further improved if panel members are instructed to rank more applicants, this needs to

be balanced with effort considerations, as well as with the genuine ability of panel members to differentiate between credible applicants who are lower on their list, an effect which is not currently captured by the model and is discussed in the next section.

Sample characteristics. As expected, the simulation of the covariate ratios (see Figure 8) indicated that the relative distribution of the covariates specifically affects the respective model coefficients, e.g. the recruitment-level covariate ratio only affected the dispersion estimates surrounding the recruitment by CV type interaction effect. Of particular interest, the observed non-linearity of this effect and its steep slope across the inspected range of values suggests that the study might face larger uncertainties surrounding the effects of variables whose distribution is highly skewed, while uncertainty levels should remain relatively similar for variables whose binomial distribution exceeds 30%. This indicates that unevenly distributed variables among early career researchers in the UK, such as gender (which varies by discipline), visa requirements, and English as a first language, can still be well estimated by the model, so long as the discrepancy is not extreme. However, this also indicates that variables with rare occurrence rates, such as applicant or panel member disability or previous training in NCV (relative to the stage of the study in the rollout of the format) will be difficult to estimate accurately. From the recruitment perspective, due to the high prevalence of positions in the STEM discipline relative to the meagre number of positions in the Arts, Humanities, and Social Sciences, the results additionally suggest it would be challenging to capture interaction effects between discipline and CV type. While our simulations were confined to binary categorical variables for simplicity, the findings should still hold for multi-level categorical variables, such that the estimation of levels which account for a small portion of the sample is expected to be poor. It should be noted that our simulations did not test for continuous covariates (e.g., academic age), which could also be included in the model.

Model assumptions

The generative model as defined in the current study carries with it several assumptions on the way the latent score is distributed and assessed by the panel members. Here we highlight some of these assumptions and discuss their ramification on the results.

During the generation of latent skill in the population, we assume that applicants come from a normal distribution, meaning we assume most are around average, while a

few are particularly good or bad. We additionally assume that this pattern persists across recruitments, i.e. should be identical across disciplines. Due to the threshold imposed by the generative model on the number of ranked candidates, only those falling at the upper part of the distribution are uniquely ranked. Given its bell shape, this means that the sparser regions of the distribution are getting ranked, where applicants' base latent skill is less dense. In other words, the applicants ranked are also those that are more distinct from one another and therefore are more likely to be similarly ranked by panel members. This suggests that the observed results might contain less certainty if the real distribution of latent skill has thicker tails (e.g., a Cauchy distribution) or is generally right skewed (i.e., fewer good candidates). Our manipulation of the distribution width approximates what we would expect to observe under these conditions, and alternative models using a different latent skill distribution might need to be fitted and contrasted with the data collected in the main phase.

A second set of assumptions pertains to the panel member's ability to accurately discern the candidate's latent skill. As described, we assume the assessment of the CV by the panel member is noisy, such that the panel member makes an observation by drawing from a normal distribution centred on the applicant's base latent skill. We likewise assume this noise does not vary between panel members, candidates, or CV types. In other words, our model assumes that all panel members are just as good at estimating the candidate's true latent skill, are just as good at estimating the latent skill of great candidates as that of poor candidates, and importantly, are as capable of accurately assessing the candidate's latent skills using a standard CV as they do using a narrative CV. While these assumptions were made for simplicity, their validity can be argued. In terms of the results, we can assume that loosening these assumptions would result in a higher level of uncertainty surrounding the model estimates, and potentially affect the results of some of the simulation experiments, particularly those related to the number of panel members, the number of credible candidates, and the max ranked candidate. Thus, the conclusions from these simulation experiments should be taken with reservations.

A third assumption our model makes pertains to the linearity of the effects on latent skill. This entails the assumption that the effect of CV type applies equally across the range of applicant latent scores, meaning that great candidates with a high latent score benefit or are disadvantaged equally by the NCV format as poor candidates with a

low latent score. Since our study design asks panel members to rank only the top applicants, meaning those whose latent score is relatively high, we have no means of testing this assumption. The closest proxy available to inform this assumption is to examine effect sizes after median splitting candidates by rank, though this would require a considerably large sample size.

Extensions and generalisation of this work

The model can easily be extended to include any number of applicant-, panel member-, and recruitment-level covariates, with the number of variables dependent on sample size. The model can also be changed to loosen the assumptions described in the previous section, and alternative models should be tested on the data collected to assess model fit, which can be fed back into the generative model for validation. It is additionally possible to include applicants and panel members as random effects in case repeated observations are made, for example if certain panel members partake in more than one recruitment, or if applicants apply to more than one position (assuming their CVs remain relatively unchanged).

The model presented here is suitable for testing other cases where data is nested within non-overlapping clusters using different raters, where raters use ranks rather than scoring, and where the researchers are interested in the estimation of group-level covariates rather than the individual-level latent score within a Bayesian framework. For example, studies testing product ranking of different products across different groups of raters where there is a covariate of interest that can be used to group products (e.g., country of origin) and interact it with a rater-related covariate (e.g., their demographics) or experiment-related covariates (e.g., mode of presentation).

Author contributions. **Noam Tal-Perry:** Data curation, Formal analysis, Methodology, Software, Validation, Writing – Original Draft, Visualisation. **Lara Abel:** Writing – Review & Editing; **Mollie Etheridge:** Investigation, Writing – Review & Editing. **Jessica Hampton:** Data curation, Investigation. **Becky Ioppolo:** Writing – Review & Editing. **Adrian Barnett:** Methodology. **Timothy R. Johnson:** Methodology, Resources, Software, Writing – Review & Editing. **Steven Wooding:** Conceptualisation, Funding Acquisition, Supervision, Writing – Review & Editing.

Competing interest. The authors have no competing interests to declare.

Grant information. This work was primarily funded by the Research England Development Fund awarded to Steven Wooding and Liz Simmonds, with additional support from several institutionally allocated sources including the Wellcome Trust Institutional Strategy Support Fund. Representatives from Research England and the Wellcome Trust are members of our External Advisory Group, along with representatives from other funding organisations who are not funding the study. The funders had no direct role in the study design, data collection or analysis.

Acknowledgements. The authors acknowledge in-kind support provided by the University of Cambridge. We would like to acknowledge Katherine Dawson contributions to the provision of resources for this study, and Liz Simmonds for the acquisition of funding. We would like to thank our Project Board: Patrick Maxwell, Chris Young, Jeremy Baumberg, Diane Coyle, Tomas Coates Ulrichsen, Tim Harper, Peter Hedges, Andrea Hudson, Steve Joy, Michael Kenny, Raphael Lyne, and Jon Simons. In addition, we would like to thank our External Advisory Group: Anne Ferguson-Smith, Steph Bales, Stephen Curry, Catherine Davies, Matthias Egger, Steven Hill, Shomari Lewis-Wilson, Molly Morgan Jones, Iben Rørbye, Tom Stafford and Karen Stroobants. Furthermore, we would like to thank our project partners: Sarah de Rijcke, Jacqui Hall, Ulrich Rößler, Sara Shinton and Martine Vernooij. We would also like to thank everyone who contributed to the study by way of participating in the job recruitments which we studied: Ali Algaddafi, Amirhossein AlizadehKhaledi, Mohd Zahid Ansari, Awais Awais, Prabhu Azhagapillai, Soumi Bairagi, Lakshmi Jaya Madhuri Bandaru, Alex Basso, Neeraj Kumar Biswas, Jenni Burt, Surajit Chakraborty, Victoria Christodoulides, Diwali Diwali, Camellia Doroody, Haolin Fei, Saptarsi Ghosh, Xiantao Hu, Peng Huang, Anit Joseph, Marlous Kamp, Muhammad Saddique Akbar Khan, Rajwali Khan, Abeedha Tu Allah Khan, Kishor Kharkwal, Benazir Khurshid, Subhrajit Konwar, Mirtunjay Kumar, Ram Kumar, August Lindemer, Rosie Lindsay, Zhe Liu, Eddison Loades, Charalambos Louca, Graham Martin, Heba Mohamed, Fizza Nazim, Van Hiep Nguyen, Rebecca Claire Oettle, Benjamin Orimolade, Ryan O'Shea, Marciano Palma do Carmo, David Pritchett, Sujeet Rai, Wei Rao, Georgios Rigas, Srinjaya Saha, Mohanapriya Saminathan, Fatemeh Sayehmiri, Subash Sharma, Artem Shushanian, Mingming Si, Paula Smith, Md Zamil Sultan, MINATI TIADI, Huseyin Unozkan, Jan van der Scheer, Raphael Luiz Vicente Fortulan, Mahak Vij, Waqas Bashir Waqas, Mohsen Yari among many other anonymous contributors. Additionally, we are grateful to the many members of the University of

Cambridge community and the Bennett Institute for Public Policy provided feedback throughout the project, particularly Simone Schnall and Andrew Rowland.

References

- Adams, E. (2021). *Narrative CVs for funding and job applications*. Rise. https://rise.articulate.com/share/NyPk_PNIENdfRS5R5catqqijzs3woS3Y#/lessons/a0G1S8sChDTr93JCgMIM5HguP42vRwqC
- Adams, E., Casci, T., Padgett, M., & Alfred, J. (2023). *Narrative CVs: Supporting applicants and review panels to value the range of contributions to research*. <https://doi.org/10.31219/osf.io/4fmj7>
- Aubert Bonn, N., Morris, J. P., Sapcaru, S., & Stroobants, K. (2024). Are Narrative CVs contributing towards shifting research culture? Workshop Report from the 2023 Recognition and Rewards Festival. *F1000Research*, 13, 332. <https://doi.org/10.12688/f1000research.146108.1>
- Bhalla, N. (2019). Strategies to improve equity in faculty hiring. *Molecular Biology of the Cell*, 30(22), 2744–2749. <https://doi.org/10.1091/mbc.E19-08-0476>
- Bordignon, F., Chaignon, L., & Egret, D. (2023). Promoting narrative CVs to improve research evaluation? A review of opinion pieces and experiments. *Research Evaluation*, 32(2), 313–320. <https://doi.org/10.1093/reseval/rvad013>
- Curry, S., de Rijcke, S., Hatch, A., Pillay, D. (Gansen), van der Weijden, I., & Wilsdon, J. (2022). *The changing role of funders in responsible research assessment: Progress, obstacles and the way ahead (RoRI Working Paper No.3)* (p. 2449096 Bytes). Research on Research Institute. <https://doi.org/10.6084/M9.FIGSHARE.13227914>

- Fritch, R., Hatch, A., Hazlett, H., & Vinkenburg, C. (2021). *Using Narrative CVs: Process Optimization and bias mitigation*. Zenodo. <https://doi.org/10.5281/zenodo.5799414>
- Gottlieb, G., Smith, S., Cole, J., & Clarke, A. (2021). *Realising Our Potential: Backing Talent and Strengthening UK Research Culture and Environment*. Russell Group. <https://realisingourpotential.russellgroup.ac.uk/#group-section-Downloads-fnkDZkcywl>
- Ioppolo, B., Hampton, J., Barnett, A., Abel, L., Etheridge, M., Tal-Perry, N., Dawson, K. M., Murray, K., Osborn, S., Wooding, S., Simmonds, L., & Matthews, Z. (2024). *Exploring the use of Résumé for Research and Innovation Narrative CVs in live postdoc recruitments*. <https://doi.org/10.17605/OSF.IO/GWA9R>
- Johnson, T. R., & Kuhn, K. M. (2013). Bayesian Thurstonian models for ranking data using JAGS. *Behavior Research Methods*, 45(3), 857–872. <https://doi.org/10.3758/s13428-012-0300-3>
- Makowski, D., Ben-Shachar, M. S., & Lüdtke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, 4(40), 1541. <https://doi.org/10.21105/joss.01541>
- Meadmore, K., Recio-Saucedo, A., Blatch-Jones, A., Church, H., Cross, A., Fackrell, K., Thomas, S., & Tremain, E. (2022). *Exploring the use of narrative CVs in the NIHR: A mixed method qualitative study*. National Institute for Health Research. <https://doi.org/10.3310/nihropenres.1115193.1>
- Pietilä, M., Kekäle, J., & Rintamäki, K. (2023, May 19). Broadening the conception of ‘what counts’ – example of a narrative CV in a university alliance. *27th International Conference on Science, Technology and Innovation Indicators (STI*

- 2023). 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). <https://doi.org/10.55835/644192caf38f9678c0feaff0>
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *DSC 2003 Working Papers*. <https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf>
- Plummer, M. (2023). *rjags: Bayesian Graphical Models using MCMC* [Computer software]. <https://CRAN.R-project.org/package=rjags>
- Posit team. (2024). *RStudio: Integrated Development Environment for R* [Computer software]. Posit Software, PBC. <http://www.posit.co/>
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Russel Group. (2021). *Research Culture and Environment Toolkit*. Russell Group. <https://russellgroup.ac.uk/media/5924/rce-toolkit-final-compressed.pdf>
- Sólymos, P. (2010). dclone: Data Cloning in R. *The R Journal*, 2(2), 29–37.
- Strinzel, M., Brown, J., Kaltenbrunner, W., de Rijcke, S., & Hill, M. (2021). Ten ways to improve academic CVs for fairer research assessment. *Humanities and Social Sciences Communications*, 8(1), Article 1. <https://doi.org/10.1057/s41599-021-00929-0>
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Thurstone, L. L. (1931). Rank order as a psycho-physical method. *Journal of Experimental Psychology*, 14(3), 187–201. <https://doi.org/10.1037/h0070025>

- UKRI. (2021). *Funders joint statement: Exploring a shared approach towards a narrative CV*. UK Research and Innovation. <https://www.ukri.org/publications/funders-joint-statement-exploring-a-shared-approach-towards-a-narrative-cv/>
- UKRI. (2023). *People and Teams Action Plan*. UKRI. <https://www.ukri.org/wp-content/uploads/2023/03/UKRI-20032023-UKRI-people-and-teams-action-plan.pdf>
- Universities UK. (2019, September). *The Concordat to Support the Career Development of Researchers*. <https://researcherdevelopmentconcordat.ac.uk/>
- Wellcome. (2022). *Wellcome Research Culture Townhalls Report*.
- Wellcome Trust. (2020). *What researchers think about the culture they work in*. <https://wellcome.org/reports/what-researchers-think-about-research-culture>
- Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1), 79–92. <https://doi.org/10.1348/000711099158973>