Rethinking the Funding Line at the Swiss National Science Foundation: Bayesian Ranking and Lottery

Rachel Heyarda,b, Manuela Otta, Georgia Salantic, and Matthias Eggerc,d

ARTICLE HISTORY
Received February 2021
Accepted May 2022

ABSTRACT
Funding agencies rely on peer review and expert panels to select the research deserving funding. Peer review has limitations, including bias against risky proposals or interdisciplinary research. The inter-rater reliability between reviewers and panels is low, particularly for proposals near the funding line. Funding agencies are also increasingly acknowledging the role of chance. The Swiss National Science Foundation (SNSF) introduced a lottery for proposals in the middle group of good but not excellent proposals. In this article, we introduce a Bayesian hierarchical model for the evaluation process. To rank the proposals, we estimate their expected ranks (ER), which incorporates both the magnitude and uncertainty of the estimated differences between proposals. A provisional funding line is defined based on ER and budget. The ER and its credible interval are used to identify proposals with similar quality and credible intervals that overlap with the provisional funding line. These proposals are entered into a lottery. We illustrate the approach for two SNSF grant schemes in career and project funding. We argue that the method could reduce bias in the evaluation process. R code, data and other materials for this article are available online.

1. Introduction
Public research funding is limited and highly competitive. Not every grant proposal can be funded, even if the idea is worthwhile and the research group highly qualified. Funding agencies face the challenge of selecting the proposals or researchers that merit support among all proposal submitted to a call for applications. Funders generally rely on expert peer review for determining which projects deserve to be funded (Harman 1998). For example, in the United Kingdom, over 95% of medical research funding was allocated based on peer review (Guthrie, Ghiga, and Wooding 2018). At the Swiss National Science Foundation (SNSF), external experts first assess the proposals which are then reviewed and discussed by the responsible panel, taking into account the external peer review (Severin et al. 2020).

The evidence base on the effectiveness of peer review is limited (Guthrie et al. 2019), but it is clear that peer review of grant proposals has several limitations. Bias against highly innovative and risky proposals is well documented (Guthrie, Ghiga, and Wooding 2018) and probably exacerbated by low success rates, potentially leading to “conservative, short-term thinking in applicants, reviewers, and funders” (Alberts et al. 2014). There is also evidence of bias against highly interdisciplinary projects. An analysis of the Australian Research Council’s Discovery Programme showed that the greater the degree of interdisciplinarity, the lower the success rate (Bromham, Dinnage, and Hua 2016). The data on gender bias are mixed, but women tend to have lower publication rates and lower success rates for high-status research awards than men (Kaatz, Gutierrez, and Carnes 2014; van der Lee and Ellemers 2015).

Several studies have shown considerable disagreement between individual peer reviewers assessing the same proposal, with kappa statistics typically well below 0.50 (Cole, Cole, and Simon 1981; Cicchetti 1993; Fogelholm et al. 2012; Guthrie et al. 2019). A similar situation is observed at the level of evaluation panels. For example, a study of the Finnish Academy compared the assessments by two expert panels reviewing the same grant proposals (Fogelholm et al. 2012). The kappa for the consolidated panel score of the two panels after the discussion was 0.23. Interestingly, the same kappa was obtained when using the mean of the scores from the external reviewers, indicating that panel discussions did not improve the evaluation’s consistency. The low inter-rater reliability means that a research proposal’s funding decision will partly depend on the peer reviewers and the responsible panel, and therefore on chance.

The wisdom of a system for research funding decisions that depends partly on chance has been questioned over many years (Cole, Cole, and Simon 1981; Mayo et al. 2006). More recently, Fang and Casadevall (2016) argued that the current grant allocation system is “in essence a lottery without the benefits of being random” and that the role of chance should be explicitly
acknowledged. They proposed a modified lottery where excellent research proposals are identified based on peer review and those funded within a given budget are selected at random. The Health Research Council of New Zealand (Liu et al. 2020), the Volkswagen Foundation (2017) and recently the Austrian Research Fund (2020) are funders that have applied lotteries, with a focus on transformative research or unconventional research ideas. The SNSF introduced a modified lottery for its junior fellowship scheme, focusing on proposals near the funding line (Adam 2019; Bieri et al. 2021).

The definition of the funding line is central in this context. Many funders rank proposals based on the review scores’ simple averages to draw a funding line (Lucy et al. 2017). Averages are understandable to all stakeholders but not optimal for ranking. Proposals around the funding line will often have similar averages to draw a funding line (Lucy et al. 2017). Averages are often weak. Kaplan, Lacetera, and Kaplan (2008) have shown that unrealistic numbers of reviewers (100 reviewers or more) per proposal would be required to detect smaller differences between average scores reliably. The same point has been made in the context of the ranking of baseball players (Berger and Deely 1988).

In this article, we show how Bayesian ranking (BR) methods (Laird and Louis 1989) can be used to define funding lines and identify applications to be entered into a modified lottery. We will describe the BR methodology with its expected rank and identify applications to be entered in the context of the ranking of baseball players (Berger and Deely 1988).

2. Methodology

2.1. A Bayesian Hierarchical Model for the Evaluation Process

Let us assume a setting where $n$ research proposals are being submitted for funding. The proposals are graded for their quality using a score on a 6-point interval scale (from 1, poor to 6, outstanding). A research proposal $i$ is evaluated by up to $m$ distinct evaluators/assessors; so that $y_{ij}$, with $i \in \{1, \ldots, n\}$, $j$ in $\{1, \ldots, m\}$, is the score given to proposal $i$ by assessor $j$.

Conditional on proposal and assessor effects, the observations $y_{ij}$ are assumed to be normally distributed with a mean that depends on the proposal and assessor. The parameters of interest for ranking are the true underlying proposal effects $\theta_i$.

\[
\begin{align*}
y_{ij} | \theta_i, \lambda_{ij} & \sim N(\bar{y} + \theta_i + \lambda_{ij}, \sigma^2) \\
\theta_i & \sim N(0, \tau_{\theta}^2) \\
\lambda_{ij} & \sim N(0, \tau_{\lambda}^2),
\end{align*}
\]

where $\bar{y}$ is the overall mean score of all proposals. The model residuals are the measurement error for assessor $j$ on proposal $i$. In the model above, $\tau_{\lambda}^2$ can be interpreted as the total variability of the proposals around the mean. To account for the tendency of assessors to be more or less strict in their scoring, we add a parameter $\lambda_{ij}$. We assume that this parameter follows a normal distribution with mean $\nu_j$ and variance $\tau_{\lambda}^2$. The assessor specific mean $\nu_j$ accounts for different scoring habits of the assessors. A stricter assessor, compared to all other assessors, will have a negative $\nu_j$; a more generous assessor will have a positive $\nu_j$.

Note that $y_{ij}$ is centered around the overall mean score only if the expected value of $\nu_j$ is 0 (which holds under our prior specification for $\nu_j$ in Section 2.3). The estimation of these parameters will be discussed later in Section 2.3.

Alternatively, a Bayesian hierarchical model for ordinal data can be used. Following the work of Johnson (2008) and Cao, Stokes, and Zhang (2010) we assume that the proposals have an underlying continuous quality trait on which the true ranking is based. The assessors essentially estimate this trait and compare it to fixed cutoffs to obtain the ordinal score. The cutoffs are assumed to be the same for each assessor. $\theta_i$ will be the underlying quality trait for proposal $i$, which is estimated by assessor $j$ to be $x_{ij}$. $x_{ij}$ is a latent, non-observed variable, while $y_{ij}$ is the observed score given by assessor $j$ using certain cutoffs $c_k$.

Let us assume that $K$ is the number of levels of the ordinal score, $k \in \{1, \ldots, K\}$, $c_0 = -\infty$ and $c_K = \infty$, then the model is defined as follows:

\[
\begin{align*}
x_{ij} & \sim N(\bar{y} + \theta_i + \lambda_{ij}, \sigma^2) \\
y_{ij} & = k \iff \lambda_{ij} \leq c_k \\
\theta_i & \sim N(0, \tau_{\theta}^2) \\
\lambda_{ij} & \sim N(\nu_j, \tau_{\lambda}^2).
\end{align*}
\]

The interpretation of the remaining parameters in model (2) stays the same as discussed for model (1).

Until now we assumed that the residual variance of the normally distributed scores is the same for all proposals $i$, namely $\sigma^2$ in Equations (1) and (2). It is possible however, that if the average score of a proposal is very low or high, then the individual scores tend to be less variable than for a proposal with an average in the middle range. The Bayesian hierarchical models presented above can be extended to heterogeneous residual variances by following the mixed-effects location scale models introduced in Hedeker, Mermelstein, and Demirtas (2008). We simply introduce proposal-specific residual variances $\sigma_i^2$ and model these variances as a function of the average scores per proposal. We also include one more effect to account for additional (i.e., beyond the variation explained by the average scores) between-proposal variation in the residual variances of the scores. More precisely, we extend the hierarchical model (1) as follows:

\[
\begin{align*}
y_{ij} | \theta_i, \lambda_{ij} & \sim N(\bar{y} + \theta_i + \lambda_{ij}, \sigma_i^2) \\
\theta_i & \sim N(0, \tau_{\theta}^2) \\
\lambda_{ij} & \sim N(0, \tau_{\lambda}^2), \\
\sigma_i^2 & = \exp(\alpha + \beta \log(\bar{y}_i) + \omega_i) \\
\omega_i & \sim N(0, \tau_{\omega}^2),
\end{align*}
\]

where $\bar{y}_i$ is the mean of the scores for proposal $i$. Note that if we assign a normal prior to $\omega_i$, then the variances $\sigma_i^2$ have a log-normal distribution (Hedeker, Mermelstein, and Demirtas 2008). The same extension can directly be applied on the ordinal outcome model in (2).

If the proposals are evaluated in different sub-panels and the results of all sub-panels have to be merged into a final ranking,
the likelihood can be adapted to include a panel effect $\delta_k$:

$$y_{ijk} | \theta_i, \lambda_{ij}, \delta_k \sim N(\tilde{y}_i + \theta_i + \lambda_{ij} + \delta_k, \sigma^2)$$  \hspace{1cm} (4)

$$\theta_i \sim N(0, \tau_\theta^2)$$

$$\lambda_{ij} \sim N(\upsilon_j, \tau_\lambda^2)$$

$$\delta_k \sim N(0, \tau_\delta^2),$$

where $k$ refers to the panel. The mean of the prior distribution on $\delta_k$ is set to 0, because we assume the panels being comparable and therefore unbiased (this assumption can of course be relaxed, for example by allowing for a panel-specific mean different from 0). The ordinal and location scale model definitions in (2) and (3) could be extended in a similar fashion to allow for different panels.

Note that in the use case where each assessor evaluates and grades only one proposal, the model would need to be simplified by deleting the assessor-specific effect and variance terms. In general, the ranking methodology described in the next sections can be applied to other Bayesian hierarchical models as long as the individual proposal quality and its uncertainty are estimated.

### 2.2. Ranking Grant Proposals

The key parameter on which we will base the ranking are the individual proposal quality and its uncertainty are estimated. The true proposal effects, respectively, the true quality traits, $\theta_i$, in the models defined above. Simply ranking proposals based on the means of the posterior distribution of the $\theta_i$’s may be misleading because it does not account for the uncertainty reflected by the posterior variances of the $\theta_i$’s. Laird and Louis (1989) introduce the expected rank (ER) as the expectation of the rank of $\theta_i$:

$$ER_i = E(\text{rank}(\theta_i)) = \sum_{k=1}^{n} \Pr(\theta_i \leq \theta_k)$$

$$= 1 + \sum_{i \neq k} \Pr(\theta_i \leq \theta_k), \text{ because } \Pr(\theta_i \leq \theta_i) = 1.$$ 

Note that ER incorporates the magnitude and the uncertainty of the estimated difference between proposals. Scaling ER to a percentage (from 0% to 100%) facilitates interpretation (van Houwelingen, Brand, and Louis 2009): PCER$_i = 100 \times (ER_i - 0.5)/n$. The percentile based on ER (PCER) is independent of the number of competing proposals and is interpreted as the probability that proposal $i$ is of worse quality compared to a randomly selected proposal. A related quantity, the surface under the cumulative ranking (SUCRA) line, summarizes all rank probabilities (Salanti, Ades, and Ioannidis 2011). It is based on the probability that proposal $i$ is ranked on the nth place, denoted by $Pr(i = n)$. Then, these probabilities can be plotted against the possible ranks, $m = 1, \ldots, n$, in a so-called rankogram. The cumulative probabilities $\text{cum}_m$ of proposal $i$ being among the $m$ best proposals are summed up to form the SUCRA of proposal $i$: SUCRA$_i = \sum_{m=1}^{n} \text{cum}_m$. The higher the SUCRA value, the higher the likelihood of a proposal being in the top ranks. As SUCRA is between 0 and 1, it can be interpreted as the fraction of competing proposals that the proposal $i$ “beats” in the ranking. As shown in Appendix B, the SUCRA can directly be transformed to ER: $ER_i = n - (n - 1) \cdot \text{SUCRA}_i$. In order to assess how reliable the votes of all the assessors were the intra-class correlation coefficient (ICC) (Shrout and Fleiss 1979) can be computed.

### 2.3. Estimation and Implementation

We pursue a fully Bayesian approach in JAGS (Just Another Gibbs Sampler) using the R JAGS interface in R (Plummer 2019). Therefore, we need to define priors on the parameters described in the models (1), (2), (3), and (4), respectively.

Since, in our case studies, the scores $y_{ij}$ all lie in the interval [1, 6], we can derive a worst-case upper bound for the variance of the $\theta_i$’s and $\lambda_{ij}$’s by assuming a uniform distribution on this interval. Apart from this upper bound, we do not have much prior information on the variability of the $\theta_i$’s and $\lambda_{ij}$’s. Therefore, as suggested by Gelman (2006), we use a uniform prior on (0, 2] for $\tau_\theta$, $\tau_\lambda$, and $\tau_\delta$ if the evaluations are done in sub-panels. Similarly, we fix a uniform prior on (0, 2] on $\sigma$:

$$\tau_\theta, \tau_\lambda, \tau_\delta, \sigma \sim U(0, 2)$$

$$\upsilon_j \sim N(0, 0.52).$$

The parameter summarizing the assessor behavior, $\upsilon_j$, will follow a normal distribution around 0, with a small variance ($\sim 0.52$), that is, about 95% of the assessors are assumed to have a assessor-specific mean in $[-1, 1]$.

For model (2) an additional prior has to be set on the cutoffs for the ordinal scale. The quantities $c_k$ denote the common cutoffs for all assessors. A uniform prior on the real line is defined on those cutoffs, with the constraint $c_k < c_{k+1}$ (Cao, Stokes, and Zhang 2010). In our case studies, we will be using $K = 6$ since the scale used in the evaluation is a six-point scale.

For model (3), we need to additionally specify hyper-priors for the parameters $\alpha$, $\beta$ and $\tau_\omega$. We assume:

$$\alpha \sim N(0, 10^2),$$

$$\beta \sim N(0, 10^2),$$

$$\tau_\omega \sim U(0, 10].$$

We implemented the presented methodology in R (see package ERforResearch available on github, the package documentation and vignette as well as the supplementary materials of this article).

### 2.3.1. Computational Aspects

Whenever MCMC methods are used, convergence diagnostics need to be considered. We decided to do visual inspections of the trace plots and computed the $\hat{R}$ values of the Gelman-Rubin convergence diagnostic (Gelman and Rubin 1992) of the most important parameters. Additionally we integrated a convergence test in the implementation, that first uses the user specified number of iterations, burn-in and adaptations of the JAGS algorithm. Then, the $R$ values of all the relevant parameters are computed. If the maximum value of those $R$ values is larger than the threshold of 1.1 recommended by Gelman et al. (2014) the latter numbers (n.burnin, n.adapt, and n.burnin) are increased until the model converges (as defined by having all $R$ values $\leq 1.1$). However, the algorithm will stop after a maximum
number of iterations (default one million) and return a warning message. More information on convergence diagnostics as well as the definition of the JAGS model and hyper-prior sensitivity can be found in Appendix C and in Sections 3 and 4 of the supplementary materials.

2.4. Strategy for Ranking Proposals and Drawing the Funding Line

To rank the competing proposals we can use the following steps:

(a) Ranking proposals based on the means of the posterior distribution of the $\theta_i$’s. This strategy incorporates the uncertainty and variation from the evaluation process, but ignores the uncertainty induced by the posterior mean’s estimation. Additionally, this is not a truly comparative ranking where all proposals are compared with every other proposal.

(b) Using the expected ranks $ER_i$ (or the equivalent PCER$_i$ and SUCRA$_i$) to compare proposals instead. This approach incorporates all sources of variation that can be quantified.

We start by plotting the ERs together with their 50% credible interval (CRI). Note that the 50% CRI was a political choice, to ensure small random selection groups (see Section 5). A provisional funding line (FL) can be defined by simply funding the best-ranked $x$ proposals. $x$ is generally determined by the available budget. This FL is then exactly defined at the ER of the last fundable proposal. The provisional funding line helps to see the bigger picture. Are there clusters of proposals? Is there a group of proposals with credible intervals of the rank that overlap with the provisional FL? The latter group should be considered for inclusion into a lottery where the proposals that can be funded are selected at random. If there is enough funding to fund all proposals in the lottery, no random selection element is needed. On the other hand, a clear distance between the ERs of non-funded proposals and the provisional FL, meaning no overlap in their 50% CRI with the provisional FL, suggests that the proposals can be easily separated with respect to their estimated quality and no random selection is needed. This strategy is what we refer to as Bayesian ranking.

3. Case Studies

In this section, we will use the approach to simulate recommendations for two funding instruments: the Postdoc.Mobility fellowship for early career researchers and the SNSF Project Funding scheme for established investigators. For all examples in the case studies we first use the continuous outcome model as described in (1) and (4), and then compare the results with the results obtained from the ordinal model described in (2). Model (3) will be used in the simulation study, see Section 4.

3.1. Postdoc.Mobility (PM) Funding Scheme

Junior researchers who recently defended their Ph.D. and wish to pursue an academic career can apply for a fellowship, which will allow them to spend two years in a research group abroad. In a first step, each proposal is allocated in one of five panels (Humanities; Social Sciences; Science, Technology, Engineering and Mathematics (STEM); Medicine; and Biology) and attributed to a referee and a co-referee who will evaluate the proposal in detail. Both of the referees then score the proposal on a six-point scale (from 1, poor, to 6, outstanding). Based on these scores, all the proposals of a panel are triaged into three groups: “fund”, “discuss” (in panel) and “reject”. The middle group’s proposals are further discussed in the respective panels to take the final funding decisions. The panel members score each of the proposals and agree on a funding line, with or without the usage of a lottery for some proposals. We used the data from the February 2020 call. Due to the pandemic, the meeting was remote, and the scoring independent: panel members did not know how fellow members scored the proposals. Table 1 summarizes the number of proposals in the different “panel discussion” groups together with the number of proposals that can still be funded and the size of the panel. Note that each panel had its own assessors, that is, panel members.

Table 1. Characteristics of panels and submissions for the February 2020 call for junior fellowships at the Swiss National Science Foundation.

| Panel    | No. of panel members | Total N. of submissions | No. of proposals discussed | No. of fundable proposals |
|----------|----------------------|-------------------------|---------------------------|--------------------------|
| Humanities | 14                   | 23                      | 11                        | 4                        |
| Social Sciences | 14           | 38                      | 18                        | 7                        |
| Biology   | 26                   | 35                      | 18                        | 8                        |
| Medicine  | 17                   | 35                      | 14                        | 7                        |
| STEM      | 29                   | 50                      | 18                        | 6                        |

Table 2. The intra-class correlation coefficients together with their 95% confidence intervals of the different panels.

| Panel    | ICC (95% CI)    |
|----------|----------------|
| Humanities | 0.33 (0.18; 0.58) |
| Social Sciences | 0.4 (0.27; 0.58) |
| Biology | 0.5 (0.37; 0.67) |
| Medicine | 0.43 (0.28; 0.63) |
| STEM | 0.31 (0.2; 0.47) |

For the funding decision, the $\theta_i$’s defined in (1) are of primary interest. We calculated the distribution of the rank of the $\theta_i$’s, and obtain the ER as the posterior mean of this distribution. Figure 1 shows the different ways of ranking the proposals, for all panels separately. The points in the left column show the ranking based on the simple averages (if two proposals have the same average grade, they are on the same rank). Next, indicated by the middle column, the proposals are ranked based on the posterior means of the $\theta_i$’s (posterior mean ranking). Finally, the points in the right column show the expected rank. A provisional funding line represented by the change of color is defined by simply funding the $x$ first ranked proposals, where $x$ is the number of fundable proposals in Table 1. This kind of presentation is solely used for illustrative purposes, and not for funding recommendations. The representation shows how the expected ranks relate to the ranking of the proposals based on their average score.
Figure 1. Junior fellowship proposals ranked based on simple averages (points on the left), the posterior means (middle) and the ER (points on the right), by the evaluation panel. Note that rank 1, displayed on the top, is the best rank. The color indicates the proposals funded if the best proposals based on the ER and constrained by the number of fundable proposals are funded.

Figure 2 plots the ERs of the same proposals together with their 50% credible intervals and the provisional funding line (= the ER of the last fundable proposal). This presentation facilitates identifying proposals that cluster around the funding line; that is, the proposals that might be included in a lottery (or random selection group). The methodology supports the following decisions:

- Humanities panel: The four best ranked proposals are funded, the remaining seven are rejected, no random selection.
- Social Sciences: The seven best ranked proposals are funded, the eleven worst ranked proposals are rejected, no random selection.
- Medicine: The five best ranked proposals are funded, the five worst ranked proposals are rejected. Two proposals are randomly selected for funding among the four proposals ranked as sixth to ninth.
- Biology: The eight best ranked proposals are funded, the ten worst ranked proposals are rejected, no random selection.
- STEM: The four best ranked proposals are funded, the ten worst ranked proposals are rejected. Two proposals are randomly selected for funding among the four proposals ranked as fifth to eighth.

Samples from the posterior distributions of all the parameters in the Bayesian hierarchical model can be extracted from the JAGS model. This allows the funders to better understand the evaluation process. As a reminder, parameter $\nu_j$ summarizes the behavior of panel member $j$. The more $\nu_j$ is negative, the stricter the scoring behavior of assessor $j$ compared to the remaining panel members. This also means that the stricter grades from assessor $j$ are corrected more, because they comply with their usual behavior. Figure 3 shows the posterior distributions of the $\nu_j$’s for the Social Sciences and Medicine panels. Note that we only present these two panels, because they are small enough to allow interpretation. The illustration of the remaining panels can be found in Figure 1.4 in the supplementary materials. In the Social Sciences, assessor 8 is a more critical panel member, whereas assessor 14 gives, on average, the highest scores. Also in the Medicine panel, the distributions for the different assessors are quite different. This illustrates how important it is to account for assessor effects.

Additionally, the posterior means of the variation of the proposal effects, $\tau^2$, for each panel with 90% CrI can be extracted:
Figure 2. Bayesian Ranking: The expected rank as point estimates (posterior mean), together with 50% credible intervals (colored boxes). The dashed blue line is the provisional funding line, that is, the ER of the last fundable proposal. The color code indicates the final group the proposal is in: accepted or rejected proposals, or random selection/lottery group.

Figure 3. Posterior distributions of the assessor-specific means $\nu_j$ in the Social Sciences and Medicine panels. Each of the 14 assessors voted on up to 18 proposals, unless they had a conflict of interest or other reasons to be absent during the vote. Note that the 14 assessors in the Social Sciences panel were not the same as the ones in the Medicine panel.
0.14 with [0.06, 0.34] for the Humanities, 0.13 with [0.07, 0.25] for the Social Sciences, 0.18 with [0.1, 0.32] for the STEM panel. The same information can be retrieved for the variation of the assessor effect with [0.09, 0.31] for the STEM panel. The same information can be retrieved for the Social Sciences, 0.18 with [0.1, 0.32] for the Biology, 0.14 with [0.06, 0.34] for the Humanities, 0.13 with [0.07, 0.25] for the Medicine panel and 0.13 with [0.34, 0.2] for the Medicine panel and 0.13 with [0.0, 0.34] for the STEM panel.

Further Figures, representing the rankograms and SUCRAs, the PCER and the ranking using the posterior mean of their 50% CrI can be found in Figures 1.5–1.7 in the supplementary material. For all the results of the panels presented above we used the convergence test described in Section 2.3.1 to ensure convergence of all chains.

Similar results are found when using the ordinal model defined in (2), see Figure 1.8 in the supplementary material. Figure 1.9 in the supplementary materials shows the Bland-Altman plot (Bland and Altman 1999) representing the agreement between the rankings based on the ordinal and continuous models, defined in Equations (1) and (2), respectively.

### 3.2. Project Funding

Project Funding is the SNSF’s most important funding instrument. Project grants support blue-sky research of the applicant’s choice. We analyzed the proposals submitted to the April 2020 Call to the Mathematics, Natural and Engineering Sciences (MINT) division. Overall, the division evaluated 353 grant proposals. The evaluation was done in four sub-panels of the same size (nine members) and a similar number of international and female members. Each panel member evaluated all proposals (unless they had a conflict of interest), and each panel defined its own funding line aiming at a similar (~ 30%) success rate.

Table 3 shows the total number of proposals, fundable proposals and average scores (from 1, poor, to 6, outstanding). Average scores were highest in panel 3 and lowest in panel 4.

Table 4 shows the ICC (with 95% confidence interval). Figure 2.1 in the supplementary materials presents all the votes of the panel members in each subpanel for each of the proposals. Compared to the previous Postdoc.Mobility case study, the intra-class correlation coefficients are higher and can be classified as good. One reason for the higher reliability here is that all Project Funding proposals are included—also the ones with low/high average scores, for which the votes are usually less variable. In contrast, for Postdoc.Mobility we only looked at the proposals with average scores in the middle range.

In this case study, we illustrate the BR with an underlying model defined in (4), where an overall ranking is aimed for, while the evaluation was done in sub-panels. Figure 4 shows the ER ordered from the best-ranked proposal (bottom left) to the worst (top right) together with their 50% credible intervals. The provisional funding line is defined as the ER of the last fundable proposal: the 106th (30% of 353) best ranked, according to its ER. Zooming in on the provisional funding line shows the cluster of proposals with similar quality and 50% credible intervals overlapping with the funding line. These proposals may be included in the lottery/random selection group.

### 4. Simulation Study

We conducted a simulation study to compare our proposed Bayesian ranking method with continuous and ordinal outcomes, homogeneous and heterogeneous residuals) to the currently used procedure, that is, the ranking based on averages. We closely follow Cao, Stokes, and Zhang (2010) to simulate the true ranks. We suppose that our panel is composed of true ranks. We therefore compare the performance of different models in a simulation study, where the true ranks are known.

As for the Postdoc.Mobility evaluations, we can also estimate the variation of the proposal, assessor and panel effects (posterior mean with 90% CrI): 1.13 with [0.99, 1.28] for \( \tau^2_j \), 0.09 with [0, 0.2] for \( \tau^2_\theta \) and 0.01 with [0, 0.12] for \( \tau^2_\tau \). A more detailed analysis of this case study can be found in the supplementary materials (especially, also the results if using model (1) on the sub-panels separately). For all the results presented above we again used the convergence test described in Section 2.3.1 to ensure convergence of all chains.

Similar results are found when using a Bayesian hierarchical model that explicitly accounts for the ordinal nature of the scores; see Figure 2.5 in the supplementary materials. Figure 2.6 in the supplementary materials shows the Bland-Altman plot; the agreement between the BR with an ordinal outcome model and the BR with a continuous outcome model. However, even though both models yield similar results, it is hard to say which modeling approach performs better without knowing exactly which proposals should have been funded or rejected. We therefore compare the performance of different models in a simulation study, where the true ranks are known.
the data, we first draw a random number \( M \in [1, 100] \) for the number of missing scores and then randomly add \( M \) missing values into the data matrix. We replicate these simulation steps \( N = 500 \) times to achieve a small enough Monte Carlo Error (see below). The code to simulate the data can be found in Appendix D. To estimate the Bayesian hierarchical models, we use 10,000 iterations, 5000 burn-in and adaptation iterations and four chains.

The true ranks are computed using the simulated true proposal effects \( (\theta_{true}^i, i = 1, \ldots, J) \): \( R_{true}^i(\theta) = \text{rank}(\theta_{true}^i) = \sum_{l=1}^J I(\theta_{true}^i > \theta_{true}^l) \). Then, in order to compare the different approaches to each other, and understand which method best estimated the true rank we will use the mean squared error (MSE) loss:

\[
\text{MSE} = \frac{1}{I} \sum_i (R_{true}^i(\theta) - R_{est}^i)^2,
\]

with \( R_{est}^i \) being the estimated rank; the rank of proposal \( i \) given by the procedure that is tested.

The approaches compared here are the ranking based on the average, the Bayesian Ranking (BR) with an underlying normal outcome model and homogeneous residuals (as described in (1)), the BR with an ordinal outcome model and homogeneous residuals (as described in (2)), the BR with a normal outcome model and heterogeneous residuals, and finally the BR with an underlying ordinal outcome model and heterogeneous residuals (both adapted from (3)). Table 5 compares the mean squared error of the different ranking approaches. Following the suggestion from Morris, White, and Crowther (2019) we additionally calculated the Monte Carlo standard error (SE) for the MSE as follows

\[
\text{Monte Carlo SE} = \sqrt{\frac{1}{N(N-1)} \sum_i ((R_{true}^i(\theta) - R_{est}^i)^2 - \text{MSE})^2}.
\]

As mentioned in Cao, Stokes, and Zhang (2010) the MSE informs us on the ability of the procedure to accurately rank all the proposals. The best performing method is the BR with underlying ordinal model and heterogeneous residuals; the model closest to the data generating process. However, the simple BR with underlying normal outcome model and homogeneous residuals outperforms the remaining approaches. Hence, applying the Occam razor principle, favoring simpler models to complex models, we would suggest to use the BR approach as defined in model (1). The reason why this simpler model performs so well is most likely due to the fact that for the ordinal outcome model (and the heterogeneous residuals) an additional latent variable, the cutoffs and other parameters have to be estimated.

### 5. Discussion

Inspired by work on ranking baseball players (Berger and Deely 1988), health care facilities (Lingsma et al. 2010) and treatment effects (Salanti, Ades, and Ioannidis 2011), we developed a Bayesian Ranking method based on a hierarchical model to support decision making on proposals submitted to the SNSF. A provisional funding line is defined based on the expected rank (ER) and the available budget. The ER and its credible interval are then used to identify proposals with similar quality and
credible intervals that overlap with the provisional funding line. These proposals are entered into a lottery to select those to be funded. The approach acknowledges that there are proposals of similar quality and merit, which cannot all be funded. Previous studies suggested that peer review has difficulties in discriminating between applications that are neither clearly competitive nor clearly non competitive (Scheiner and Bouchie 2013; Fang and Casadevall 2016; Klaus and Alamo 2018). Decisions on these proposals typically lead to lengthy panel discussions, with an increased risk of biased decision making. The method proposed here avoids such discussions and thus may increase the efficiency of the process, reduce bias and costs.

The Bayesian ranking methodology compares every assessor to every other panel member. The ER considers all uncertainty in the evaluation process that can be observed and quantified. It accommodates the fact that different assessors have different grading habits. The model can also be adjusted for potential confounding variables, such as external peer reviewers’ characteristics that influence their scores. A recent analysis of 38,250 peer review reports on 12,294 SNSF project grant applications across all disciplines showed that male reviewers, and reviewers from outside Switzerland, awarded higher scores than female reviewers and Swiss reviewers (Severin et al. 2020). Jayasinghe, Marsh, and Bond (2003) modeled peer review records from the Australian Research Council using a multilevel cross-classified model in order to identify researcher and assessor attributes that influence the evaluation. They found low reliability in the scores attributed by the assessors, no gender effects, but an effect of the status of the University and the academic rank of the first applicant.

We agree with Goldstein and Spiegelhalter (1996) who argued that “no amount of fancy statistical footwork will overcome basic inadequacies in either the appropriateness or the integrity of the data collected.” In an ideal world, all proposals would be evaluated by as many experts it takes to ensure that meaningful differences between aggregated scores can be detected with confidence. Evaluations would be unbiased and describe nothing else but the quality of the proposals. Human nature and limited resources regarding time and funding sadly prevent this ideal situation from becoming a reality. The evaluation of grant proposals will always be subjective to some extent and affected by unconscious biases and chance. However, we are confident that the method presented here is an improvement over the commonly used approaches to ranking proposals and defining funding lines. Our approach should not be seen as a mechanistic cookbook approach to decision making but as a tool that can provide decision support for proposals of similar or indistinguishable quality around the funding line. For example, judgement continues to be required to decide whether a lottery should be used or not.

We applied the approach to two instruments in career and project funding at the SNSF. Our case studies addressed the specific context of the SNSF and the two funding schemes and results may not be generalizable to other instruments or funders. Further, we acknowledge that the team carrying out this study included several researchers affiliated with the SNSF. As the researchers’ expectations might influence interpretation, critical comment and review of our approach from independent scholars and other funders will be particularly welcome.

In our first model, we treated the ordinal scores (from 1 to 6) as continuous variables, thus, assuming that the distance between each set of subsequent scores is equal. This assumption might not always be appropriate, but it builds on the currently used average-based ranking method, as this procedure also assumes normality of the grades. We showed how our initial model can be generalized to allow for an ordinal outcome variable. However, because of the fact that for the ordinal model extra parameters have to be estimated and in the simulation studies this more general model did not always perform better, we suggest to use the simpler model based on the normal likelihood.

The choice of priors in Bayesian models is always disputable. Especially for the variance parameters, alternative prior distributions could be investigated, like half-normal and half-Cauchy priors, whereas inverse-Gamma priors are generally not recommended (Gelman 2006). Instead of a Gibbs sampler, the brms-package in R (Bürkner 2018) can be used, which makes the programming language Stan accessible with commonly used mixed model syntax, similar to the lme4-package by Bates et al. (2015). Another approach is to use the R code provided by Lingsma et al. (2010). They implemented a frequentist approach where the posterior means and variances of the proposal-specific random intercept are approximated. However, if the proposal effects are a posteriori dependent, because of the same assessors evaluating the same set of proposals, the Bayesian approach is easier to implement.

The use of a lottery to allocate research funding is controversial. At the SNSF the applicants are informed about the possible use of random selection, thus, complying with the San Francisco Declaration on Research Assessment (DORA 2019), which states that funders must be explicit about assessment criteria. Of note, in the context of the Explorer Grant scheme of the New Zealand Health Research Council, Liu et al. (2020) recently reported that most applicants agreed with random selection. So far, the SNSF received no negative or positive reactions to the use of random selection from applicants (Bieri et al. 2021). The lottery element being controversial is also the reason why we start using the Bayesian ranking methodology with a 50% CrI. This ensures that the random selection groups are reasonably small.

An alternative to ranks based on simple averages of scores is the procedure employed by the National Institutes of Health (NIH). The NIH uses percentile rankings to define a funding line1: proposals are first discussed and assigned scores from 1 to 9 by all panel members. For each proposal, all scores are then averaged to obtain an overall impact score. Finally, percentiles are determined by matching overall impact scores against historical relative rankings (based on the last three calls/review cycles). The NIH procedure makes a relative ranking possible by putting the different scores in context, however, when drawing the funding line (or payline), the uncertainty of the estimated quantities is not taken into account. This is an important difference to a ranking based on PCER (which is also percentile-based), ER or SUCRA. Ignoring uncertainty is especially problematic since there is evidence that the percentile

---

1see https://www.niaid.nih.gov/grants-contracts/understand-paylines-percentiles.

---
scores employed by the NIH are poorly predictive of grant productivity (Fang, Bowen, and Casadevall 2016).

In conclusion, we propose that a Bayesian modeling approach to ranking proposals combined with a modified lottery can improve the evaluation of grant and fellowship proposals. More research on the limitations inherent in peer review and grant evaluation is needed. Funders should be creative when investigating the merit of different evaluation strategies (Severin and Egger 2020). We encourage other funders to conduct studies and test evaluation approaches to improve the evidence base for rational and fair research funding.

Appendix

Appendix A: Analytical Formulas

This section will give some insights on how the previously discussed quantities can be computed analytically rather than using MCMC samples. The probability of \( \theta_i \) being smaller than \( \theta_k \), which is used for the ER, can be computed as follows:

\[
\text{Pr}(\theta_i \leq \theta_k) = \text{Pr}(\theta_i - \theta_k \leq 0) = \Phi \left( \frac{\hat{\theta}_k - \hat{\theta}_i}{\sqrt{\text{var}(\hat{\theta}_i) + \text{var}(\hat{\theta}_k) - 2 \text{cov}(\hat{\theta}_i, \hat{\theta}_k)}} \right),
\]

where \( \Phi \) is the standard normal cumulative distribution function. Here, \( \hat{\theta}_i \) denotes the posterior expectation of \( \theta_i \) and \( \text{var}(\hat{\theta}_i) \) and \( \text{cov}(\hat{\theta}_i, \hat{\theta}_k) \) the corresponding posterior variance and covariance. If the \( \theta_i \)'s are a posteriori independent (which does in general not hold for model (1)), then \( \text{cov}(\hat{\theta}_i, \hat{\theta}_k) = 0 \) for \( i \neq k \).

There is also an analytical version of the posterior variance of the rank \( R_i \) discussed in the Appendix of Laird and Louis (1989) that can be used to compute confidence intervals rather than credible intervals. According to the latter authors, this posterior variance is given by:

\[
\text{var}(R_i) = \text{var} \left( \sum_{k=1}^{n} 1(\theta_i \leq \theta_k) \right) = \sum_{k=1}^{n} \text{var} \left( 1(\theta_i \leq \theta_k) \right) + 2 \sum_{k=1}^{n} \sum_{l=k+1}^{n} \text{cov}(1(\theta_i \leq \theta_k), 1(\theta_i \leq \theta_l)) = \sum_{k=1}^{n} \text{Pr}(\theta_i \leq \theta_k) \cdot (1 - \text{Pr}(\theta_i \leq \theta_k)) + 2 \cdot \sum_{k=1}^{n} \sum_{l=k+1}^{n} \left[ \text{Pr}(\theta_i \leq \min(\theta_k, \theta_l)) - \text{Pr}(\theta_i \leq \theta_k) \cdot \text{Pr}(\theta_i \leq \theta_l) \right].
\]

Appendix B: Relationship between ER and SUCRA

In the following, the relationship between the SUCRA and the ER is derived. Note that the ER, that is, the expectation of the rank, can be expressed in terms of the rank probabilities as follows: \( ER_i = \sum_{j=1}^{n} \text{Pr}(i = j) \).

\[
\begin{align*}
SUCRA_i &= \frac{1}{n-1} \sum_{m=1}^{n-1} \text{cum}_{im} = \frac{1}{n-1} \sum_{m=1}^{n-1} \sum_{j=1}^{m} \text{Pr}(i = j) \\
(n - 1)SUCRA_i &= \sum_{m=1}^{n-1} \sum_{j=1}^{m} \text{Pr}(i = j) \\
(n - 1)SUCRA_i - n &= \sum_{m=1}^{n-1} \sum_{j=1}^{m} \text{Pr}(i = j) - \sum_{m=1}^{n} \sum_{j=1}^{m} \text{Pr}(i = j),
\end{align*}
\]

because \( \sum_{j=1}^{n} \text{Pr}(i = j) = 1 \)

\[
\begin{align*}
-(n - 1)SUCRA_i + n &= -\sum_{m=1}^{n-1} \sum_{j=1}^{m} \text{Pr}(i = j) + \sum_{m=1}^{n} \sum_{j=1}^{m} \text{Pr}(i = j) \\
n - (n - 1)SUCRA_i &= \sum_{m=1}^{n} \sum_{j=1}^{m} \text{Pr}(i = j) \\
n - (n - 1)SUCRA_i &= \sum_{j=1}^{n} \text{Pr}(i = j) + \sum_{j=2}^{n} \sum_{j=1}^{j-1} \text{Pr}(i = j) + \cdots + \sum_{j=n}^{n} \text{Pr}(i = j) \\
(n - 1)SUCRA_i &= \sum_{j=1}^{n} \text{Pr}(i = j) = \text{ER}_i
\end{align*}
\]

Appendix C: Implementation of the Bayesian Model in \texttt{jags} and Convergence Diagnostics

The following code describes the definition of the model in \texttt{R} through the package \texttt{rjags}. Note that this JAGS model definition refers to the model described in (4), exactly as it is used in Section 3.2. Find more alternative model definitions in the function \texttt{get_default_jags_model()} in our package \texttt{ERforResearch}.

```r
# Likelihood:
for (i in 1:n) {  # i is not the application
  if (i > 1) {  # i is not the application
    review_grade[i] ~ dnorm(mu[i], inv_sigma2)  
  } else {  # i is the application
    review_grade[i] ~ dnorm(mu[i], inv_sigma2)  
  }
  review_grade[i] <- overall_mean + application_intercept[i] + section_intercept[num_section[i]] + voter_intercept[j, l]  
}

# inv_sigma2 is precision (1 / variance)
mu[i] <- overall_mean + application_intercept[i] + section_intercept[num_section[i]] + voter_intercept[j, l]  

# Priors:
rank_theta[1:n_application] <- rank(-application_intercept[])  

# Ranks:
for (j in 1:n_application) {  # j is the application
  application_intercept[j] ~ dnorm(0, inv_taus_application2)  
}

for (l in 1:n_voters) {  # l is the voter
  voter_intercept[j, l] ~ dnorm(nu[l], inv_taus_voter2)  
}
```
for (l in 1:n_voters){
  nu[l] ~ dnorm(0, 4)
}
for (l in 1:n_section){
  section_intercept[l] ~ dnorm(0, inv_tau_section2)
}
sigma ~ dunif(0.000001, 2)
inv_sigma2 <- pow(sigma, -2)
inv_tau_application2 <- pow(tau_application, -2)
tau_application ~ dunif(0.000001, 2)
inv_tau_voter2 <- pow(tau_voter, -2)
tau_voter ~ dunif(0.000001, 2)
inv_tau_section2 <- pow(tau_section, -2)
tau_section ~ dunif(0.000001, 2)
}\n
Trace plots and the $\hat{R}$ values of the Gelman-Rubin convergence diagnostic of the most important parameters can be computed; see Figure 3.1 and Table 3.1 of the supplementary materials. For reproducibility purposes, the length of burn-in and adaptations phases, and the final number of iterations used in our case studies can be retrieved from our supplementary materials as well as the seed used for the sampling. With this information all models converged (with $\hat{R}$ values lower to 1.1). Note that for convenience, we kept the numbers of iterations high and did not try to optimize. The supplementary materials also briefly discusses the sensitivity of the results to the choice of the hyperpriors; see Section 4 of the supplementary materials.

Appendix D: Code Simulating Data

Find the R code used for simulating the data in Section 4.

# Simulate data after Cao et al.:
set.seed(0307)
J <- 10
j <- 1:J
I <- 50
N <- 500
sims <- list()
for (n in 1:N){
  alpha <- runif(J, -.5, .5)
sigma <- runif(J - 1, 0, .5)
sigma <- c(sigma, .9)
c <- rep(NA, 5)
c[1] <- -1
c[2] <- -.5
c[3] <- 0
c[4] <- .5
c[5] <- 1
theta <- rnorm(I, 0, 1)
x <- y <- matrix(NA, nrow = I, ncol = J)
for (ii in 1:I){
  for (jj in 1:J){
    x[ii, jj] <- rnorm(1, alpha[jj] + theta[ii], sigma[jj])**2)
    y[ii, jj] <- 1
    if (x[ii, jj] > c[1]) & (x[ii, jj] <= c[2])
      y[ii, jj] <- 2
    if (x[ii, jj] > c[2]) & (x[ii, jj] <= c[3])
      y[ii, jj] <- 3
    if (x[ii, jj] > c[3]) & (x[ii, jj] <= c[4])
      y[ii, jj] <- 4
  }
}
# How many NAs/COIs do we want to introduce?
n_nas <- sample(0:100, 1)
# Which elements of the matrix will be NA?
# Make sure, we did not set a whole column or row to NA
while (test){
  where_nas <- sample(1:(I*J), n_nas)
  xx <- x; yy <- y
  xx[where_nas] <- NA
  yy[where_nas] <- NA
  test <- any(rowSums(is.na(xx)) == J) | any(colSums(is.na(xx)) == I)
}
# Creation of data in list format
sims[[n]] <- list(x = xx, y = yy, theta = theta, rank = rank(-theta))

Supplemental Materials

An online fully reproducible supplement is provided which uses an R (ERforResearch) package with the implementation of the above presented methodology (see snsf-data.github.io/ERpaper-online-supplement/). The data used in the case studies can be downloaded from Zenodo: https://doi.org/10.5281/zenodo.4531160.

Acknowledgments

We are grateful to Hans van Houwelingen and Ewout Steyerberg for helpful comments on an earlier version of this manuscript and to the National Research Council of the SNSF for fruitful discussions. We also thank Malgorzata Roos for further feedback on the manuscript as well as on the implementation in R.

References

Adam, D. (2019), “Science Funders Gamble on Grant Lotteries,” Nature, 575, 574–575. doi:10.1038/d41586-019-03572-7. [111]
Alberts, B., Kirschner, M. W., Tilghman, S., and Varmus, H. (2014), “Rescuing US Biomedical Research from its Systemic Flaws,” Proceedings of the National Academy of Sciences of the United States of America, 111, 5773–5777. doi:10.1073/pnas.1404402111. [110]
Austrian Research Fund. (2020), “1000 Ideas Programme,” Available at https://www.fwf.ac.at/en/research-funding/fwf-programmes/1000-ideas-programme. [111]
Bates, D., Maechler, M., Bolker, B. M., and Walker, S. C. (2015), “Fitting Linear Mixed-Effects Models Using lme4,” Journal of Statistical Software, 67, 1–48. doi:10.18637/jss.v067.i01. [118]
Berger, J. and Deely, J. (1988), “A Bayesian-Approach to Ranking and Selection of Related Means with Alternatives to Analysis-of-Variance Methodology,” Journal of the American Statistical Association, 83, 364–373. doi:10.2307/2288851. [111,117]
Bieri, M., Roser, K., Heyard, R., and Egger, M. (2021), “Face-to-face Panel Meetings Versus Remote Evaluation of Fellowship Applications: Simulation Study at the Swiss National Science Foundation,” BMJ Open, 11, e047386. doi:10.1136/bmjopen-2020-047386. [111,118]
Bland, J. M., and Altman, D. G. (1999), “Measuring Agreement in Method Comparison Studies,” Statistical Methods in Medical Research, 8, 135–160. doi:10.1177/096228029900800204. [116]
