SNSF Report

Chiara Gilardi, Leslie O'Bray, Carla Schärer and Tommaso Portaluri

5 April 2018

Introduction

The Swiss National Science Foundation (SNF) is a research funding agency which disseminates yearly, on behalf of the Swiss Government, billions of CHF to the best researchers in Switzerland. This report contains a statistical analysis performed on three data sets provided by SNF, containing information on the applications for funding received in 2016, the corresponding and the scores given by both internal and external evaluators.

The analysis performed for SNF had a three-fold aim, corresponding to the following three research questions: 1) Is gender bias occurring at any stage of the SNSF evaluation process? Is the gender of the main applicant influencing the rating of the application? 2) To what extent the different steps of the evaluation and the different criteria within each step determine the final funding decision? 3) When an application is approved, but the budget requested is cut, how can we explain this?

The SNSF evaluation procedure is a multi-step process (involving external reviewers, internal referees, and an internal board) which takes into consideration both the track record of the applicant and the quality of the project (see Appendix for a more detailed description of the evaluation procedure).

Several studies (Witteman et al., 2017; Solans-Domenech et al., 2017) have shown that female applicants' projects get higher score when the application is blinded. Moreover, female applicants receive usually higher grades for projects and lower grades for track record. Hence, after investigating the gender dimension to identify possible biases in the evaluation procedure, the focus of the analysis will be the relative importance of of the criteria for funding (applicant's track record vs. quality of the proposal) and, also, of each step of the evaluation procedure (which opinion is more likely to determine the final decision – the external referee's or the board's?). Possible interactions between the gender dimension and the second research question will also be investigated (for instance, by taking into account also the gender of evaluator or the percentage of female referees).

Data Description

We have three data sets: Applications, External Reviewers and Internal Referees. They contain respectively information about the SNSF project funding applications, the evaluation of the applications by external peer reviewers and the evaluation of the proposals by external the internal referee and co-referee (when available). For a full description of the data & variables, please see the Appendix.

Cleaning the Data

We decide to work with only complete applications, i.e. project for which we have information from all the three data sets.

To avoid a temporal trend, we are only considering application from 2016.

In both the external and internal step, we encountered applications which had several reviews per application. For the sake of our analysis, in these scenarios we computed the mean grade for each criteria, so that each application had a "single" score for each criteria assessed on. In doing so we also introduced a new variable, PercentFemale, which calculated the percent of female reviewers out of all reviewers of a single application (ranging from 0 to 1).

All applications with a grade were converted to an ordinal factor.

Specific to each data set, this are the detailed considerations:

Applications

We decide to consider only the MainDiscipline2 because for MainDiscipline we have 118 levels, while for the other only 21.

There is one application for which we do not know the gender of the applicant, and therefore we decided to omit that observation from the analysis.

We will also not consider the variables "CallTitle", "Professorship", "AcademicAge". The first one, because we consider it has nothing to add to the model. The two last, due to the fact that there are a considerable number of NA's on those variables (around 93% of the observations).

```
##
##    Assistant professor with tenure track
##                                      102
## Assistant professor without tenure track
##                                       54
##                       Associate professor
##                                      237
##                           Full professor
##                                      512
##  Honorary professor or Titular professor
##                                       77
##                                     None
##                                      430
##                     Professor at UAS / UTE
##                                       74
##                         Visiting professor
##                                        4
##                                     <NA>
##                                    20037
```

```
## [1] 0.9307846
```

```
## [1] 0.9361267
```

External Reviewers

Reviewers always have the option to choose not to consider or to give the grade "0" when reviewing an application. Some might be mistakes, in others cases there might be a conflict of interest, or they might be very ambivalent about the project. Therefore, we did not considered observations with this grades.

One of the questions evaluated in the applications is "Broader impact (forms part of the assessment of scientific relevance, originality and topicality)". For the time frame we are considering, in all the applications this grade was NA. Hence, we omit this variable from our model.

- **ProposalCombined**: We created a new variable to summarize the assessment of the scientific proposal in the external review step. This is a simple mean of the grade given for Suitability and Scientific Relevance. This helped to isolate the effect of the grade given to the scientific proposal, versus the applicant track record, as well as to ensure easier comparison with the internal review step.
- **PercentFemale**: As previously mentioned, we introduced a new variable calculating the percent of reviewers of each application that is female.

Internal Referees

There were 22 observations (1 for the time frame we are dealing with) for which only demographic information was available, no grades were given. We decide to omit those observations.

Also we decide to not consider the Referee role as a variable in our model, as the majority of the evaluations has only one referee.

- **PercentFemale**: As previously mentioned, we introduced a new variable calculating the percent of reviewers of each application that is female.

```
##
##          Applicant Explicit inclusion            Recusal
##                  2                  20                  8
##            Referee     Second referee                <NA>
##              15766                 870                1276
```

Exploratory Analysis

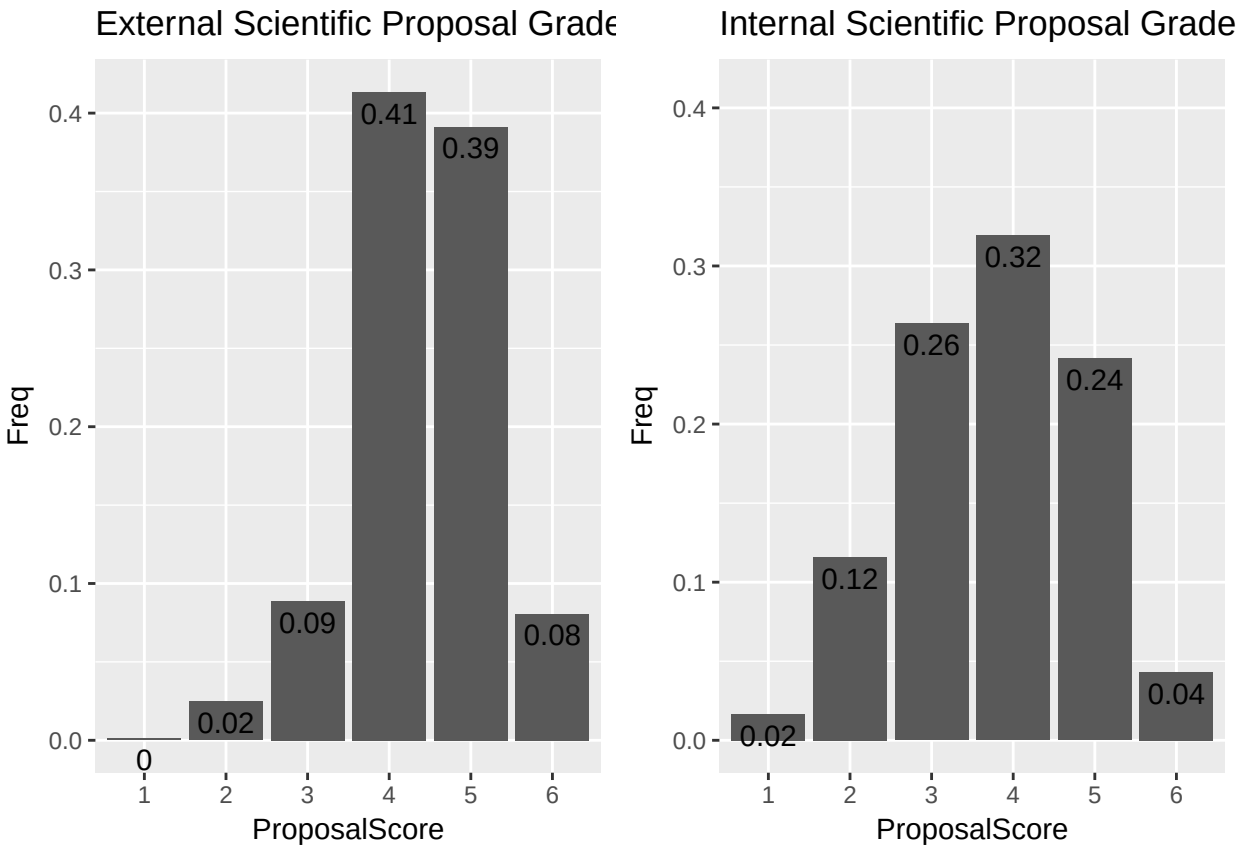
In our exploratory analysis, we discovered a few interesting insights, that relates to the findings we will discuss from our analysis.

Distribution of Grades between the External & Internal Review Step

Since the external & internal step both assess candidates on the same criteria (the strength of the scientific proposal, and the strength of the applicant), on the same ordinal scale (from poor to outstanding), we were interested to see if the distribution of grades are the same. We would expect different distributions for the Overall Grade vs the Ranking, since those have two different measurements, however we were interested to see if for the same absolute ranking, the external and internal reviewers had different perspectives on the application. After combining the Suitability & Scientific Relevance grades given to a candidate in the external review step, we can compare the average grade given for the Scientific Proposal in the two steps, as well as the grade given for the Applicant Track Record in both steps.

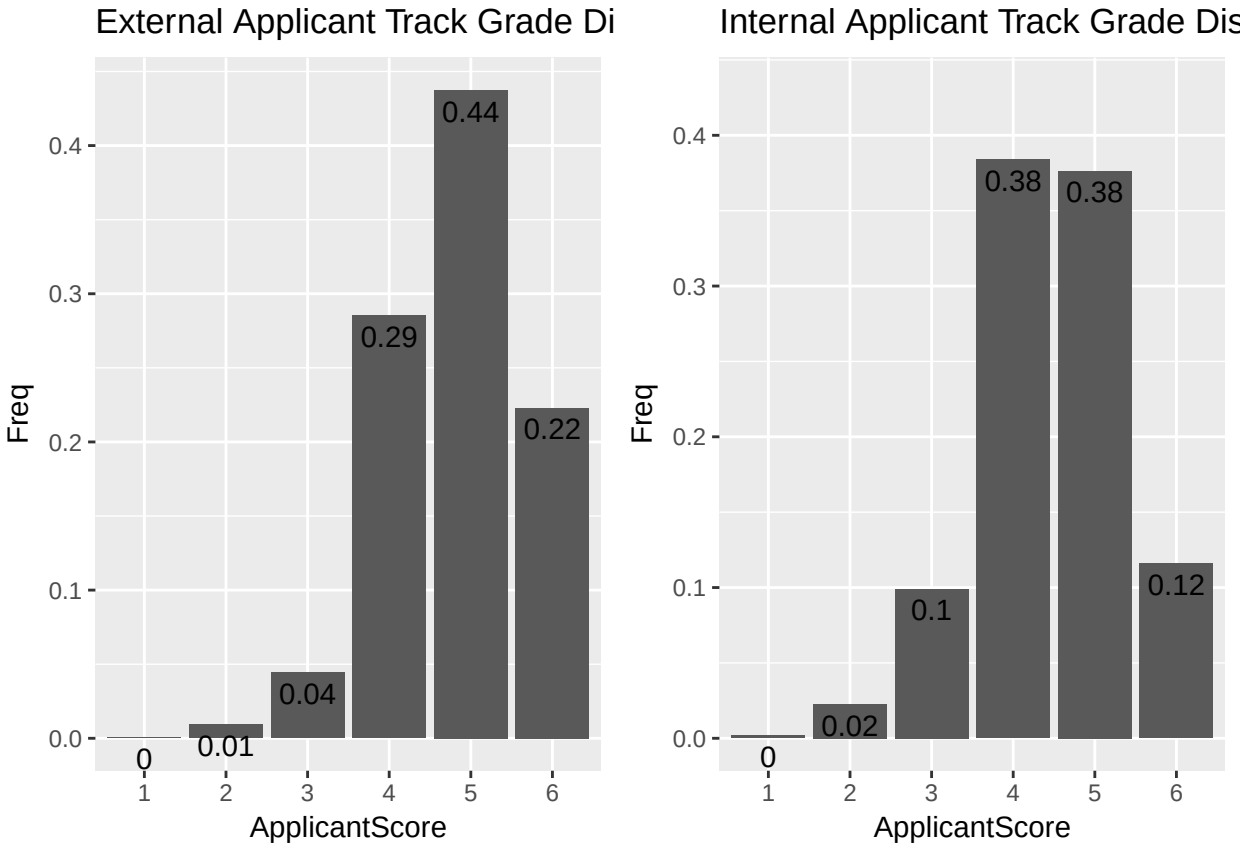
We see that the External Reviewers are more generous with their grades; for the strength of the Scientific Proposal, 48% of proposals are considered “excellent” or “outstanding”, versus only 28% in the internal review step.

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 -- 0.45
```



Similarly we see the same pattern with Applicant Track Record: 66% of Applicant Track records are considered "excellent" or "outstanding" by the External Reviewers, versus merely 50% by the Internal Reviewers.

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 -- 0.45
```



Since we noticed this discrepancy, we wanted to quantify how differently the grades were to one another. To assess the agreement between the two steps, for the same criteria, we used Cohen's Kappa. Cohen's Kappa measures the proportion of agreement between two raters assessing something on an ordinal scale, accounting for the fact that there will always be some proportion by random chance. An important specification of Cohen's Kappa is the weight given to the measurements. If the external & internal reviewers both assessed the Applicant Track Record as "excellent", that would be considered full agreement. However, we want to allocate partial credit if the rating is a level close to it. We used a linear weight up to distance 2, and after that gave no credit. (In this example, if one rater gave an "outstanding" or "very good", that would be considered a distance of one and be weighted by 0.8. If the second rater assessed the Applicant Track to be "good", which is a distance of two away from excellent, that would be weighted as 0.6). Anything with a distance of 3 or more (in this example, if the second rater gave a rating of "average"), we allocated no weight, as the difference between average and excellent is quite large.

From this, we found that there was just moderate agreement between the two steps when using the weighted kappa, for both the grades given for the Scientific Proposal & the Applicant Track Record.

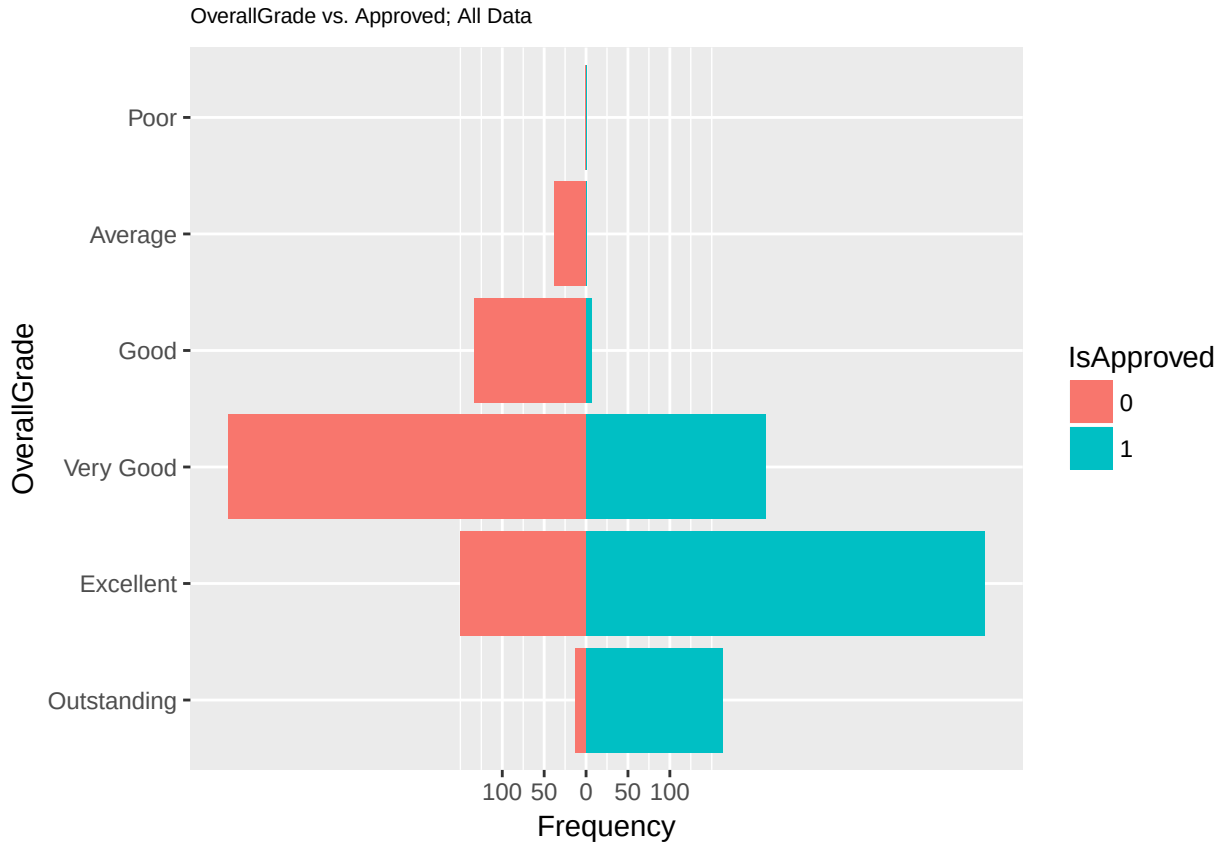
```
## [1] "Cohen's Kappa for Applicant Track Record"

## Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)
##
## Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries
##                  lower estimate upper
## unweighted kappa  0.21     0.24  0.28
## weighted kappa    0.35     0.42  0.49
##
##  Number of subjects = 1623

## [1] "Cohen's Kappa for Scientific Proposal"
```

```
## Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)
##
## Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries
##                  lower estimate upper
## unweighted kappa  0.21     0.24  0.28
## weighted kappa    0.35     0.42  0.49
##
##  Number of subjects = 1623
```

Impact of Internal Reviewers on Funding

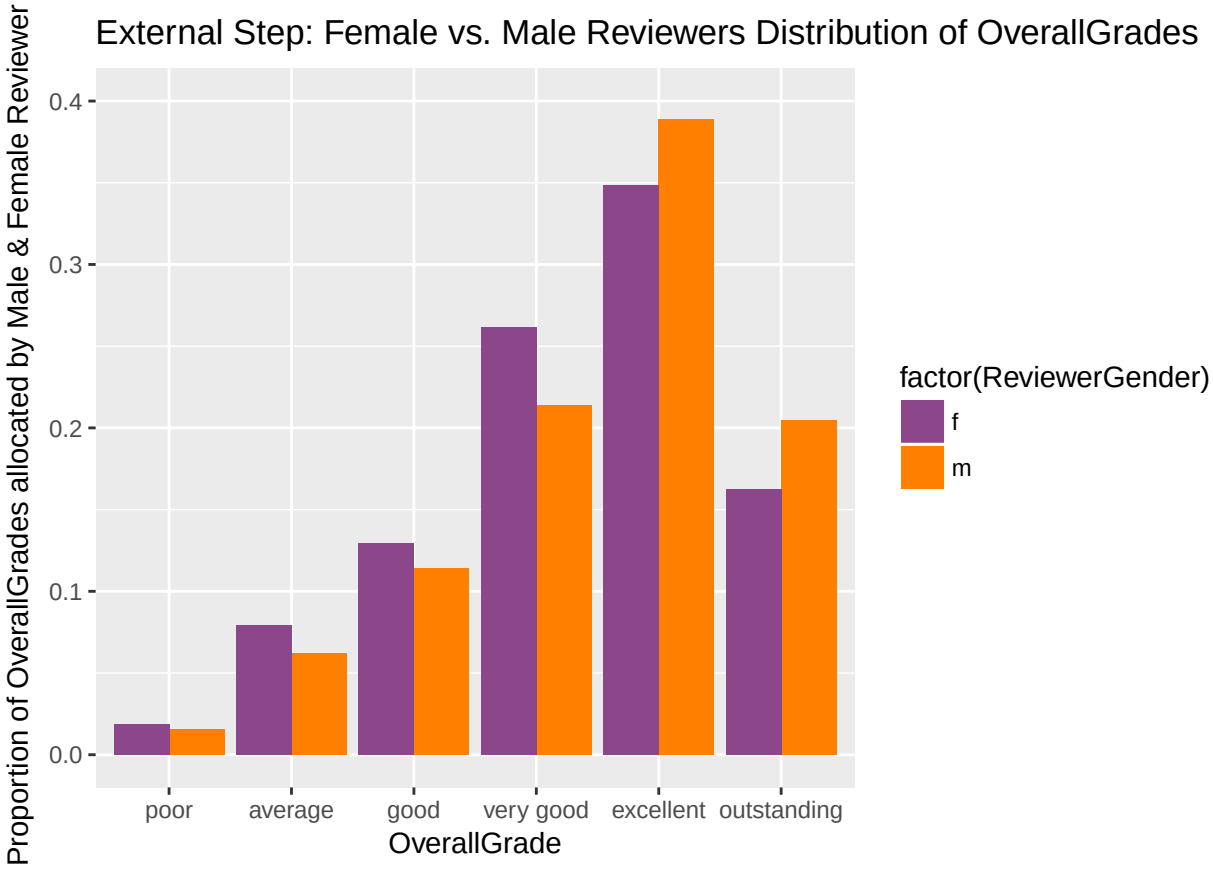
We wanted to understand if this discrepancy between grades had an impact on whether an application is funded. To do this, we visualized the summary grade given to an application, and whether that application is funded or not. As we can see here, there are several applications with an OverallGrade of "excellent" or "outstanding" that end up not approved. It highlights that not only do the Internal Reviewers give tougher grades in general than the external step, but they also consider some "excellent" and "outstanding" applications by the external reviewers to be not of the quality that deserves funding. This trend is true in all divisions and both genders, please refer to the appendix to see the specific graphic.

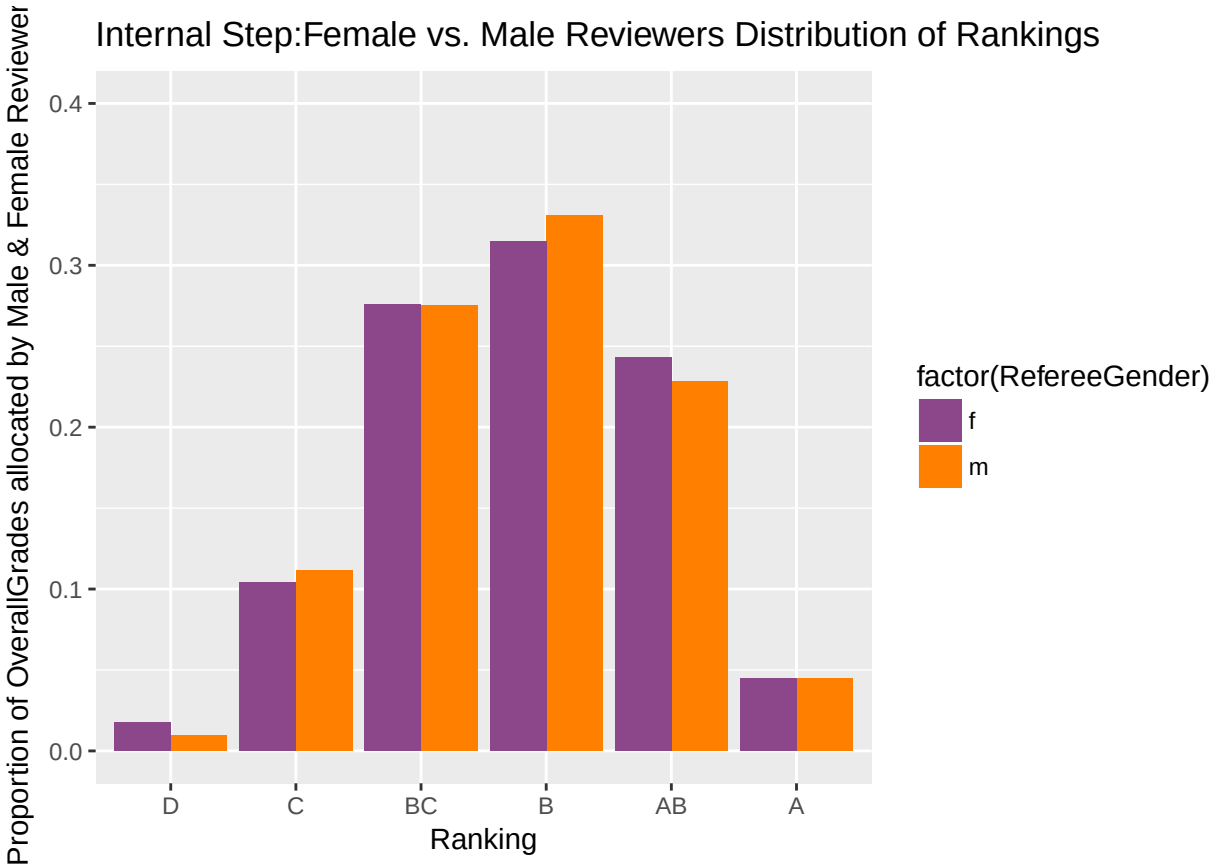
Our conclusion for this is that the internal step is very consequential, and the difference in the rating they give translates into differences in whether an application gets funded or not.



Distribution of Grades by the Gender of the Reviewer

The third interesting insight we found was when we investigated the impact of the gender of the person reviewing the data. We look at the relative frequencies of grades given by male and female reviewers, to applicants, regardless of gender. Within the external step in particular, we found that female reviewers give proportionally fewer "excellent" and "outstanding" grades, compared to their male counterparts. Within the internal step, we did not notice a particular difference, though we will consider the impact of the gender of the reviewer more rigorously in our analysis.





Gender Bias

Analysis

Results

Relative Importance of the Different Steps

Our second research question was to assess the relative importance of each step in the process, and the relative importance of each criteria within each step.

Analysis

Most Important Step

To approach the question of which step in the process is most important, we first fit a logistic regression with IsApproved as our binary response variable using the glm function in R. We fit a full model with all potential demographic predictors and interactions across the different steps of the process, and the summary grade given to an application in the external (OverallGrade) and internal (Ranking) step. To address the first part of the question (relative importance of each step in the process), we used only the summary grade in each step due to the correlation between the individual grades given within each step and the summary grade given. With the full model (predictors: Gender, Division, Age, IsContinuation, InstType, log(AmountRequested),

PercentFemale, Ranking, OverallGrade, Gender:Division, PercentFemale:Gender), we achieved a pseudo-R^2 value of 0.7251, indicating that percent of the variation in Y can be explained by the model.

As our goal was to explain the most important factors, we then did backwards variable selection using the AIC. This left us with a model with only 4 predictors: Ranking, OverallGrade, Age, and IsContinuation. The pseudo R^2 measure of this model is 0.7234, which indicates that this simplified model nearly explains exactly as much variance in the data as the full model, and so we can be content to use just the small model.

Now that we've reduced our model to 4 predictors, we wanted to understand exactly how important each of those predictors are to the final funding decision. To do this, we performed permutation tests on the variables. For each of these 4 predictors, we randomly permuted the values of that predictor and refit the the model. We did this 1000 times for each predictor, and calculated the average change in pseudo R^2 when we permuted that particular variable. If permuting a variable changes the pseudo r^2 a lot, this means that that variable was an important predictor in our regression.

The results from this permutation test were the following: permuting Ranking decreased the pseudo R^2 on average by 0.3028. Since the pseudo R^2 metric is between 0 and 1, this is a big difference. Permuting the OverallGrade decreased the pseudo R^2 on average by 0.0078, showing that the Ranking is much more important than the OverallGrade. Permuting Age and IsContinuation had very small impact, with an average decrease of 0.0021 and 0.0012 respectively. From this regression, and the subsequent permutation test, we can conclude that the Internal Step is by far the most important step in the process.

Most Important Criteria Within Each Step

The second aspect of this question was to identify what was the most important criteria within each step. To understand this, we again did a permutation test of the different predictors determining the summary grade given in each the external and the internal review step. We used the Ordinal Regressions from earlier: one for the external OverallGrade using the demographic data, Scientific Proposal grade, and Applicant Track grade as predictors, and a second one predicting the Ranking using the demographic data, Scientific Proposal grade, and Applicant Track grade as predictors.

In each of these two regressions, we again computed the variable importance by permuting the values of the predictors one at a time. Since it is an Ordinal Regression, and there is no R^2 equivalent to measure the goodness of fit, we assessed goodness of fit based on the percent of variation in the AIC, another measure of goodness of fit. For both the external and the internal step, we found permuting the grade given to the Scientific Proposal by far had the biggest impact on the quality of the regression. This led us to conclude that the grade given to the Scientific Proposal far outweighs the grade given to the Applicant Track Record, or any of the demographic predictors, in explaining the overall grade given to an application.

Results

Budget Cuts

Analysis

Results

Conclusion

Appendix

Detailed Data Description

Applications

- **AmountRequested**: Rounded to the next 10k CHF
- **AmountGranted**: Rounded to the next 10k CHF
- **IsApproved**: 1 if the application was approved, 0 if it was rejected
- **GradeFinal**: Comparative ranking of the application as determined by the evaluation body (the division of the National Research Council). A: “belongs to the 10% best percent”; AB: “10% are worse, 75% are better”; B: “50% are worse, 25% are better”; BC: " 25% are worse, 50% are better“; C”10% are worse, 75% are better“; D:”90% of the applications are better"
- **Division**: Evaluation Body in which the application was evaluated. Division 1 evaluates Social Sciences and Humanities; Division 2 Mathematics, Natural Sciences and Engineering; Division 3 Biology and Medicine
- **MainDiscipline**: as chosen by the applicant from the SNSF discipline list
- **MainDisciplineLevel2**: category in the SNF discipline list grouping disciplines into fields of research
- **CallTitle**: Call for proposals under which the application was submitted. Applications from the same Call are evaluated together, i.e. in competition to each other
- **CallEndDate**: Submission deadline of the Call
- **ResponsibleApplicantAcademicAgeAtSubmission**: Years since the applicant’s PhD at time of submission; data only available since mid 2016
- **ResponsibleApplicantAgeAtSubmission**: Biological age of the applicant at time of submission; data only available since mid 2016
- **ResponsibleApplicantProfessorshipType**: employment situation of the applicant at time of submission; data only available since mid 2016
- **Gender**: of the main applicant
- **NationalityIsoCode**: Nationality of the main applicant
- **IsHasPreviousProjectRequested**: 0 if it is the applicant’s first application at the SNSF, 1 if not
- **InstType**: Type of institution where the applicant is employed
- **IsContinuation**: 1 if the project is a thematic continuation of a previously approved project, 0 if not

- **ProjectID**: Anonymized identifier of the application

Referee Grades

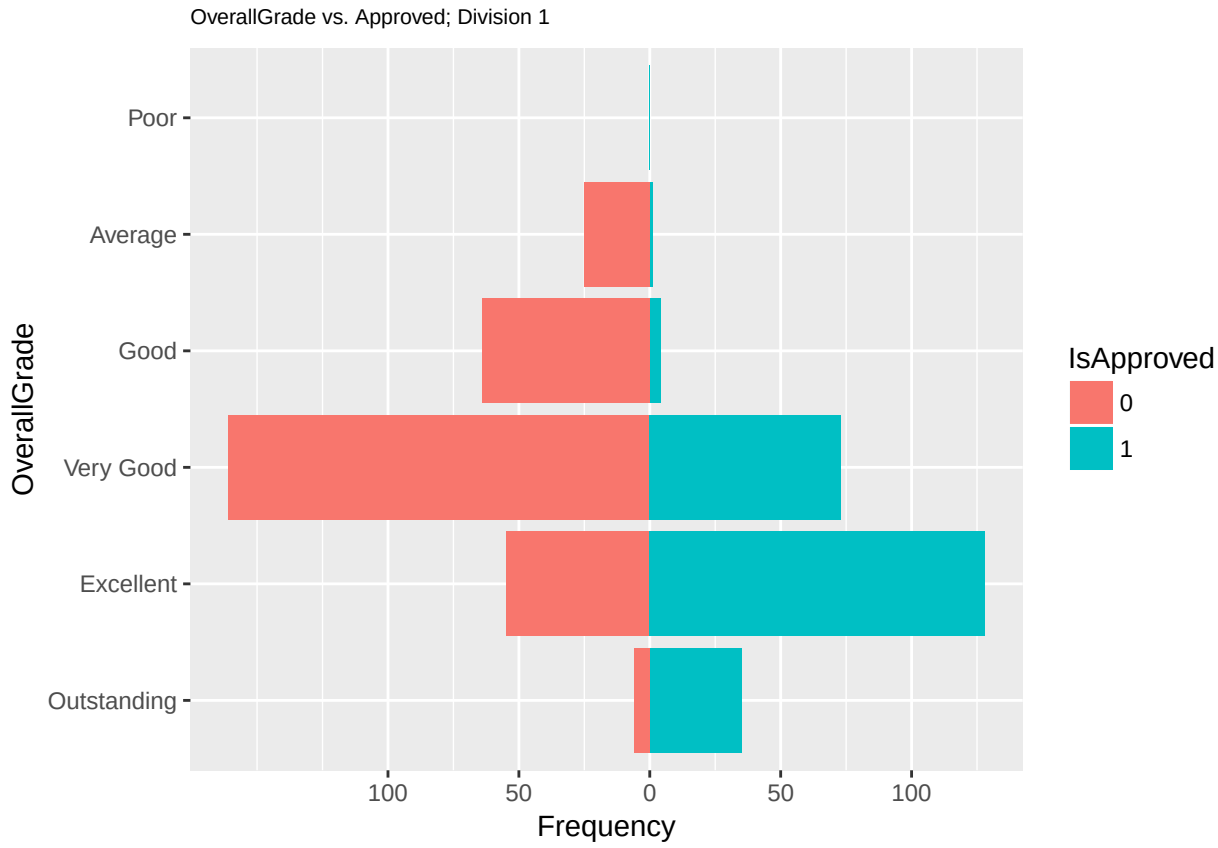
- **Question**: Evaluation criterion
- **QuestionRating**: The (co-)referee's assessment of the evaluation criterion
- **OverallRanking**: The (co-)referee's overall comparative ranking of the application. A: "belongs to the 10% best percent"; same scale as the GradeFinal
- **RefereeRole**: Some applications have one referee evaluation, some have two. The role indicates who was the primary and who was the secondary referee (also called co-referee)
- **RefereeGender**
- **IDs**: Anonymized identifiers of the application, the referee and the evaluation by the referee

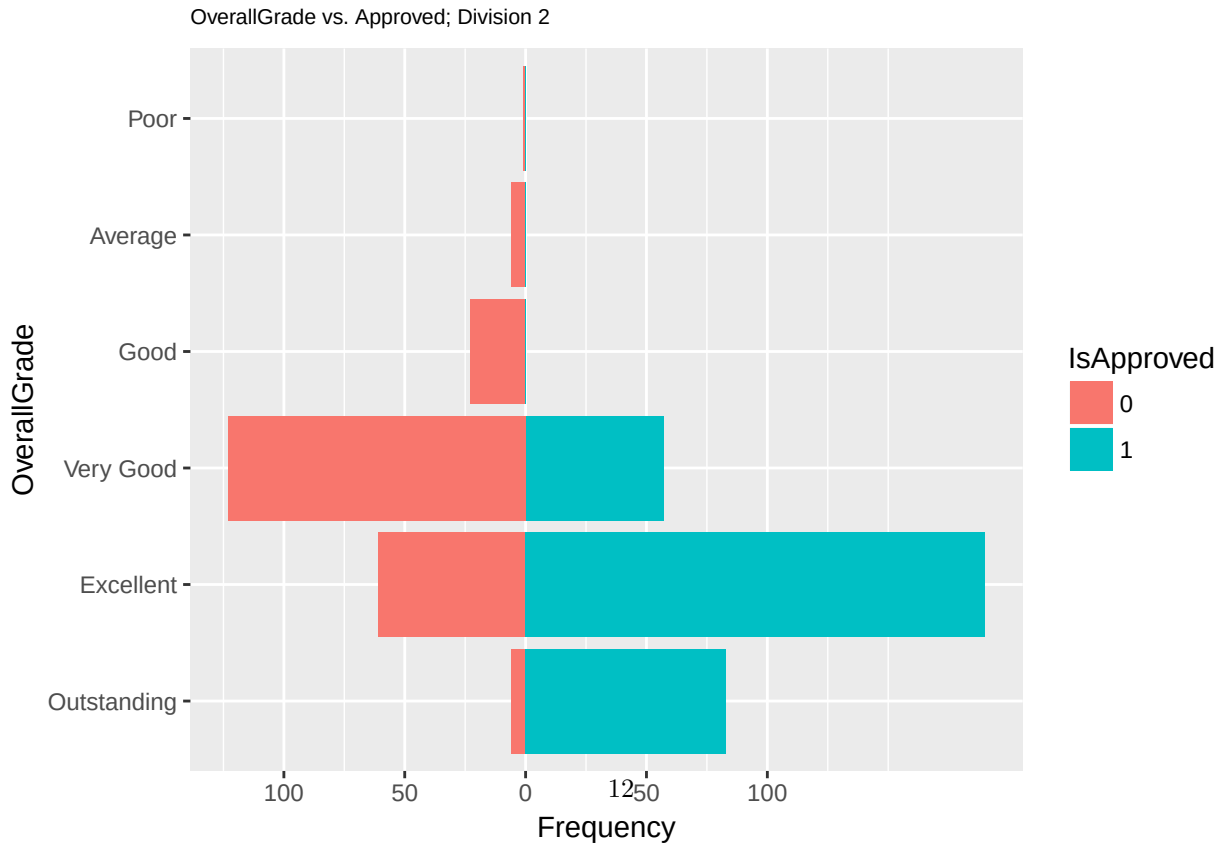
Reviews

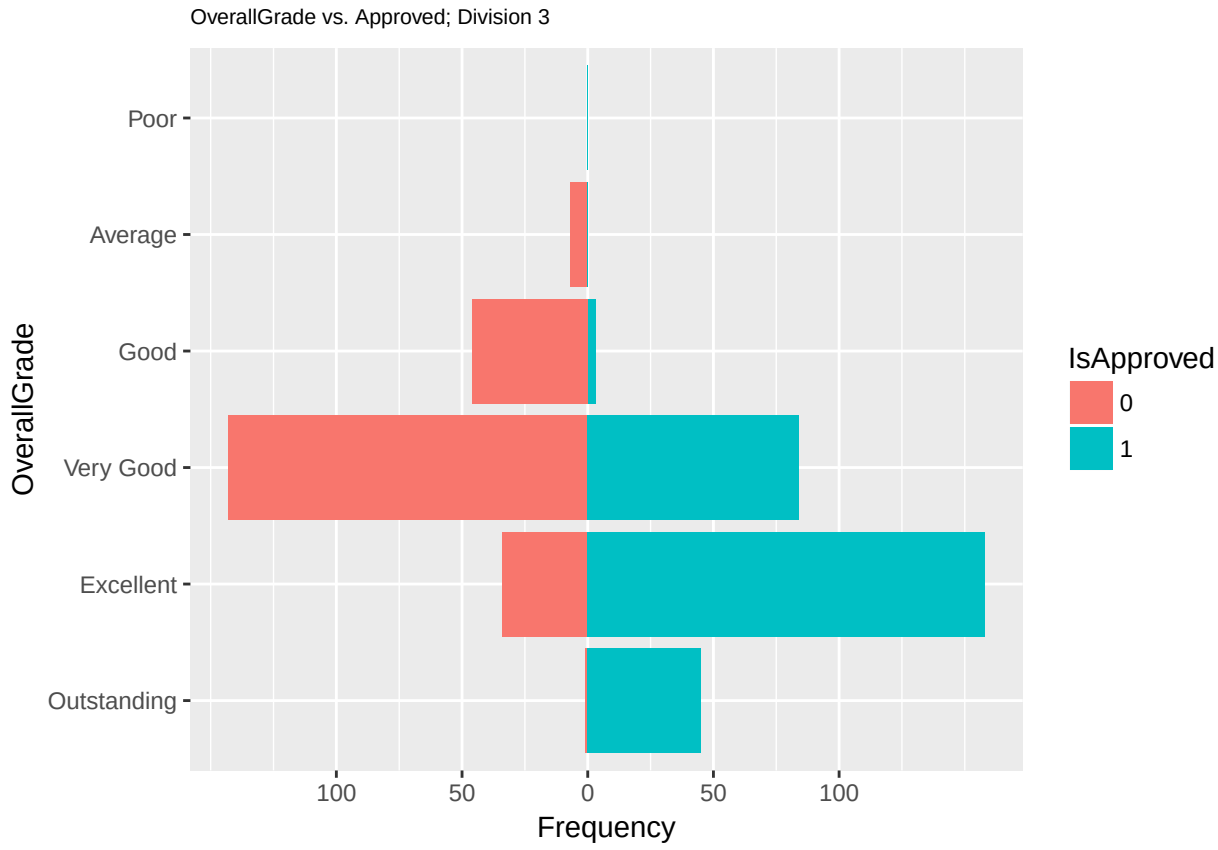
- **Question**: Evaluation criterion
- **QuestionRating**: The external reviewer's assessment of the evaluation criterion
- **OverallGrade**: The external reviewer's overall assessment of the application
- **SourcePerson**: Who suggested the reviewer?
- **Gender**
- **Country**: where the reviewer is located. Not always known
- **EmailEnding**: ending of the reviewer's email address. Might be used as an approximation of the country where the reviewer is located in cases where this data is missing
- **IDs**: Anonymized identifiers of the application, the reviewer and the review

Exploratory Analysis

OverallGrade vs. IsApproved, by Division






OverallGrade vs. Approved; Division 3
OverallGrade
Poor
Average
Good
Very Good
Excellent
Outstanding
100
50
0
50
100
Frequency
IsApproved
0
1

OverallGrade vs. IsApproved, by Gender