Select the Winners Fast

Haitao Wang
Chunfeng Huang
Hongling Rao
Center for Astrophysics
University of Science and Technology of China
Hefei, China

Advisor: Qingjuan Yu

Summary

Assuming that judges are ideal, we provide a model to determine the top W papers in almost the shortest time. We use a matrix record the orderings that we get from judges, and we reject as many papers as possible after each round.

We then consider real-life judges and estimate the probability that the final W papers contain a paper not among the best $2W$ papers.

Furthermore, considering the possibility of systematic bias in a scoring scheme, we improve the model by using a Bayesian estimation method, which makes it possible to some extent to compare different judges' scores.

We performed many computer simulations to test the feasibility of our model. We find that our model would be improved by increasing the number of papers selected from the first round. We also made a stability analysis by altering P, J, W and got an empirical formula to predicts the total time of judging.

We used data from real life to test our model and got a perfect result: For $P = 50$, $J = 3$, and $W = 2$, we get the first- and third-best papers with our scheme; with $W = 3$, we got the top three papers.

We conclude by summarizing a practicable and flexible scheme and offering some suggestions, and estimating the budget with an empirical formula.

Assumptions

- The judges are equal. None is more authoritative than the others.
- When a judge is evaluating a paper, the judging result is not influenced by adventitious factors, such as taking bribes.
- The time that a judge takes is proportional to the number of papers to read.

- There exists an objective criterion with which we can tell which of two papers is "better." Therefore, we can use an absolute rank-order or absolute scores to describe the quality of the papers measured by the criterion.
- The absolute rank-order is transitive: If A is better than B and B is better than C, we can say A is better than C.

Analysis of Problem

Our primary goal is to include the top W papers among the "best" $2W$ papers.

A subsidiary goal is that each judge read the fewest possible number of papers. We interpret this goal into two points:

- It is the duration of the whole judging process, the total time for all rounds, that is constrained by funding. If the time for a round is how long it takes the judge who has the most papers to read, it is wise to distribute the papers to the judges as evenly as possible in each round.
- We want to get as much information as possible.

The two usual methods of judging are rank-ordering and numerical scoring. Systematic bias is possible in a scoring scheme, that is, each judge may have a subjective tendency in scoring, which results in incomparability among scores given by different judges. However, it is reasonable to believe that the scores that the same judge gives to different papers are comparable, even if they are obtained in different rounds. Therefore, compared with a rank-ordering method, scoring is a more meaningful way to record the results for papers judged in earlier rounds. We use a scoring scheme instead of a rank-ordering scheme in our later model, so a paper need not be read more than once by the same judge. Note that we do not compare the scores of different judges directly, that is, we mainly use scores to obtain a rank-ordering.

We first consider the simplified problem with the significant assumption that the ordering from each judge's evaluation coincides with the absolute ordering. In this event, we can definitely find the best W papers. Furthermore, we can optimally adjust the allocation of papers in every round to get an efficient scheme.

But judges in real life cannot rate the papers with perfect precision. For example, a paper with absolute rank 7 (we denote it $P_{(7)}$) may get a higher score than $P_{(6)}$ from a judge. We call that *misjudgment*. Misjudgments prevent us from getting the best W papers, so their effect must be taken into account.

There are also subjective differences among the scorings of different judges. For example, for the same two papers, one judge may give 80 and 83, while another gives 65 and 72. If we know the distribution of each judge's scores, we can to some extent compare scores given by different judges. The real distribution for each judge is unknown, so we have to use estimates.

Table 1.
Notation.

Symbol	Meaning
P	total number of papers
W	number of winners
J	number of judges
T	total judging time (or number of papers that can be judged in the time)
P_i	paper i
$P_{(i)}$	paper with the absolute rank of i
S_i	the absolute score for paper i
$P_i > P_j$	paper i is better than paper j in absolute rank-order
$P_i(A) > P_j(A)$	paper i is better than paper j in judge A's opinion
R_i	number of papers currently known to be better than P_i
ORD	matrix of currently known relations between pairs of papers
$\lceil x \rceil$	the smallest integer not less than x
$N(\mu_0, \sigma_0^2)$	normal distribution with mean μ_0 and standard deviation σ_0
σ_1	standard deviation of the judges' scoring
μ_j, σ_j	mean and standard error of judge j's scoring
$\hat{\mu}_j, \hat{\sigma}_j$	estimated values of μ_j, σ_j
P_{error}	probability of error occurring

Design of the Model

Top W in the Least Time

Ideally, the ordering in each judge's opinion coincides with the absolute ordering, expressed mathematically by

$$P_i(A) > P_j(A) \Leftrightarrow P_i > P_j.$$

So, based on the transitivity of the absolute score, if $P_i(A) > P_j(A)$ and $P_j(B) > Pk_(B)$, we can say that $P_i > P_k$.

To find the top W as soon as possible, as many papers as possible should be rejected after each round. So if there are W papers or more in the current paper pool that are better than P_i , reject P_i.

In the first round, P papers are dispatched to J judges evenly. After performing the above rejection rule, each judge selects W papers. In later rounds, how do we dispatch the remaining $W \cdot J$ papers to judges to obtain the greatest number of new orderings from each round?

Let us consider the simple case in which $W = 2$, with $P_1 > P_2$ and $P_3 > P_4$ known after the first round. If we compare P_2 with P_4 (or P_1 with P_3), no matter what the result is, we can always gain an extra relation (if, say, $P_2 > P_4$, the extra relation is $P_1 > P_4$). If we compare P_1 with P_4 (or P_2 with P_3), on some occasions no extra relations can be obtained. A similar result holds for $W = 3$.

This example indicates a fact: If we use *current rank* R_i to denote the number of papers known to be better than P_i from current known information, we should try to distribute the papers with close current rank to the same judge in order to get more relations in a round.

Use matrix ORD to describe known orders. Define ORD_{ij} by

$$\text{ORD}_{ij} = \begin{cases} 1, & \text{if } P_i > P_j; \\ -1, & \text{if } P_j > P_i; \\ 0, & \text{if } P_j = P_i; \\ \infty, & \text{if } P_i \text{ and } P_j \text{ have not been compared by any judge.} \end{cases}$$

At the beginning of a round, we dispatch papers to judges and judges give each paper a score. We then find every P_i and P_j that have scores from the same judge in the finished rounds and fill ORD_{ij} and ORD_{ji}. We replace ORD with its transitive closure [Wang 1986], which, put simply, adds all the indirect order gained from ORD_{ij} into the matrix ORD. At the end of each round, for each paper P_i, calculate R_i from ORD and reject P_i if $R_i \geq W$. Repeat the above process until the final W papers are left.

Consider Misjudgment

By *misjudgment*, we mean that the final W papers are not the best W. If the final W papers contain a paper not among the best $2W$, an *error* occurs.

Assume that for a paper with an absolute score of μ_1, the score given by a certain judge is a random number following a normal distribution $N(\mu_1, \sigma_1^2)$. The standard deviation σ_1 is the parameter that describes the degree of precision in a measurement. Misjudgment originates from the deviation of a judge's scoring from the absolute score, as **Figure 1** shows. The shaded area in **Figure 1** shows the misjudgment area.

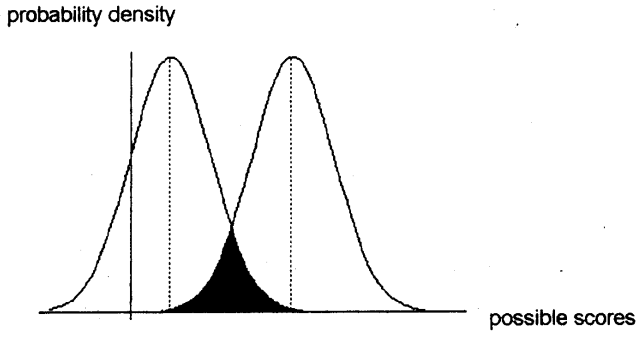

Figure 1. Possible score distributions for two papers.

There must be a distribution of the absolute scores of all the papers. We assume that it is a normal distribution $N(\mu_0, \sigma_0^2)$, so that the ratio σ_1/σ_0 reflects the judge's ability to distinguish the quality of these papers and also determines the probability of misjudgment. Using the basic model, given P, J, W, and σ_1/σ_0, we can estimate the probability of error (P_{error}). If the probability is small enough, we can expect the model to provide the desired result.

Taking the random feature of scoring into account, some conflict is likely to happen, such as $\text{ORD}_{ij} = 1$ but judge A's scores show $P_j(A) > P_i(A)$. One

way to solve the conflict is to find all judges who have read both P_i and P_j, sum up the scores given to P_i and P_j by these judges, and determine a new ORD_{ij} by comparing the two sums.

Systematic Bias Among Judges

Considering differences among the scoring tendencies of different judges (systematic biases), it is undesirable that each judge select out the same number of papers in the first round, for then it will be more likely that excellent papers will be rejected in the first round.

Instead, when the first round of judging is over, we input the scores of each group of papers into computer, which gives the estimate of each judge's parameters (mean score and standard deviation) and computes each group of papers' score threshold for rejecting papers corresponding to a certain absolute level. This way, excellent papers have less possibility of being rejected in the first round. Estimating the parameters of all the judges enables us to compare the scores from different judges to some extent.

We use Bayesian estimation [Box and Tiao 1973] to determine the estimate of judge j's parameters (μ_j, σ_j). Suppose that judge j gives scores $S_1, \ldots, S_n$ to papers $P_1, \ldots, P_n$. We use the method of maximum likelihood to estimate σ_j:

$$\hat{\sigma}_j = \frac{1}{n} \sum [S - E(S_i)]^2.$$

We then use Bayes's method to estimate μ_j. In reality, we may have a priori knowledge of each judge's scoring tendency. Even if not, we still have reason to assume an a priori distribution of each judge's score. If the prior parameters are (μ_0, σ_0^2), then the posterior parameter is

$$\hat{\mu}_j = \frac{n \cdot E(S)}{n + \left(\frac{\hat{\sigma}_j}{\sigma_0}\right)^2} + \frac{\left(\frac{\hat{\sigma}_j}{\sigma_0}\right)^2 \cdot \mu_0}{n + \left(\frac{\hat{\sigma}_j}{\sigma_0}\right)^2}.$$

Then we can use $\text{quantile}(N(\hat{\mu}_j, \hat{\sigma}_j^2), \text{LEVEL})$ as the score threshold, where $1-\text{LEVEL}$ is the expected proportion of papers should be retained. One suitable value is

$$\text{LEVEL} = 1 - \frac{W \cdot J}{P}.$$

Test of the Model

The most important test is to verify that the model makes sense. We do a computer simulation to see how our model behaves as the two practical factors are gradually taken into account. As detailed later, our results agree with our expectations (**Feasibility Test**). In addition, a finding in the course of testing

leads us to make some improvement to the model. A more thorough test is made by revising the parameters (**Stability Test**). Lastly, we try to apply our model to more complicated example in real world. The results meet reality very well (**A Real-Life Example**).

Feasibility Test

We fix $P = 100$, $J = 8$, and $W = 3$.

Test of Basic Model

We assign P papers absolute scores of $1, 2, \ldots, 100$. These values are used only to provide the relative order of the papers.

All papers are randomly allocated to 8 judges at the beginning of the simulation. We calculate the total judging time.

The results of 1,000 iterations of simulation (see **Figure 2**) show that the basic model can select the top three in quite a short time, as we analyzed before.

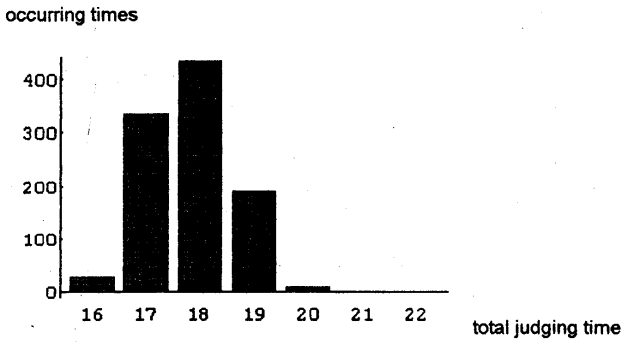

Figure 2. Frequency vs. total judging time.

Take Misjudgment into Account

These two simulations are vital:

- *Simulating the distribution of the absolute score.* Generally, we have reason to use a normal distribution. In order to assign the scores of 100 papers, we generate 100 random numbers following $N(60, 30^2)$, truncated at 0 and 100.
- *Simulating the score given to one paper.* We simulate a judge's scoring by adding a normal random number to the absolute score of the paper.

The quantity σ_1/σ_0 should be fairly small (say, ≤ 0.1), because a judge should have good competence in judgment. We take σ_1/σ_0 = $1/30$, $2/30$, and $3/30$ as cases in our simulation.

We also did a theoretical estimate for these cases of P_{error} under the worst of circumstances. Take $\sigma_1/\sigma_0 = 3/30$ for example. An error occurs when one or

more of $P_{(7)}, P_{(8)}, \ldots$ enter the final three. The probability of $P_{(7)}$ entering the final three contributes the most to P_{error}. We let $P_{\text{error}}(i,j)$ be the probability of misjudging papers i and j. We approximate P_{error} by the probability of $P_{(7)}$ entering the final three:

$$\begin{aligned} P_{\text{error}} &\approx \frac{1}{8}P_{\text{error}}(3,7) + \left(\frac{1}{8}\right)^2 \sum P_{\text{error}}(3,i)P_{\text{error}}(i,7) + \cdots \\ &\approx \frac{1}{8}P_{\text{error}}(3,7) + \left(\frac{1}{8}\right)^2 \sum_{i=4}^{6} P_{\text{error}}(3,i)P_{\text{error}}(i,7). \end{aligned}$$

We computed $P_{\text{error}}(i,j)$ using Mathematica by calculating the area of the shaded region in **Figure 1**. In this way, we get the estimate $P_{\text{error}} \approx 0.4\%$.

The results of the simulations accord with the theoretical estimate (see **Table 2**).

Table 2.

Results of 1,000 trials for each value of σ_1/σ_0 vs. theoretical estimates.

$P = 100$, $J = 8$, $W = 3$

σ_1/σ_0	Mean T	Max T	Errors	Observed P_{error}	Estimate of P_{error}
1/30	17.7	21	0	.000	10^{-7}
2/30	17.7	21	0	.000	.0006
3/30	17.8	21	4	.004	.004

An Extra Improvement to the Model

The simulation results demonstrate that the model behaves reasonably so far. Surprisingly, a slight modification improves the model remarkably. If in the first round we select more papers, say W_1 instead of W, and select W papers from the next round, we find that P_{error} declines greatly but the total judging time is scarcely affected. The chances of a excellent paper being rejected in the first round are much more than in the later round, because the papers rejected after the first round are read by only one judge, while those rejected later are read by more judges. **Table 3** gives simulation results for several values of W_1.

Table 3.

Results of 1,000 trials for each value of W_1.

$P = 100$, $J = 8$, $W = 3$

W_1	Mean T	Max T	Errors
6	17.9 ± 0.9	20	0
5	17.8 ± 0.8	20	0
3	17.8 ± 0.8	20	1

Take Judges' Systematic Biases into Account

We might as well simulate the scores from different judges by using the normal distribution with randomly generated mean value and variance.

Using the method offered in **Design of Model**, we get the results of **Table 4**. With increasing LEVEL, the total judging time declines but more errors occur; it is difficult to minimize both time and number of errors.

Table 4.

Results of 1,000 trials for each value of LEVEL.

$P = 100$, $J = 8$, $W = 3$

LEVEL	Mean T	Max T	Errors
50%	22.0	24	0
70%	19.5	23	3
75%	18.7	21	4
80%	18.0	21	4
85%	17.2	20	7
90%	16.3	19	11

Stability Test

We change the parameters P, J, W to test the model's stability. **Table 5** gives the results of 100 iterations for each of several groups of parameters.

Table 5.

Results of 100 trials, for $\sigma_1/\sigma_0 = 3/30$, for each combination of values of P, J, and W.

P	J	W	Mean T	Max T	Errors	$\lceil P/J \rceil + W + 2$
50	4	4	18.1 ± 1.2	22	0	19
80	8	3	14.9 ± 0.9	17	0	15
100	7	3	19.5 ± 0.7	21	0	20
	8	3	17.7 ± 0.8	20	0	18
	8	4	18.8 ± 0.8	21	0	19
		5	19.8 ± 0.9	22	1	20
	10	3	15.4 ± 0.9	18	1	15
120	8	3	19.8 ± 0.9	22	0	20
140	8	3	23.0 ± 1.0	27	1	23
	13	1	14.4 ± 0.7	16	3	14
		2	15.8 ± 0.8	18	1	15
		3	16.9 ± 0.8	19	0	16
		5	18.2 ± 0.9	21	0	18

Analyzing these data, we discover an empirical formula

$$\left\lceil \frac{P}{J} \right\rceil + W + 2,$$

which fits the data for the average value of T wonderfully. Another finding is that small W/P will cause considerable P_{error}. So when W/P is too small (say $\leq 1/100$), the model does not work well. But properly reducing the number of papers rejected in each round will reduce P_{error}.

A Real-Life Example

One idea for testing our model would be to use data from a tennis competition. The table of international standings can be treated as the absolute order, and the result of each formal match acts as a "judge." It is a pity that we have no data!

So we use a substitute for the data of a real competition. We obtained from our department real scores of 50 students for three semesters taking the same three courses. We consider that the sum of each student's three scores stands for the student's level in this major; we take it as an absolute score, and this gives the absolute ordering. In any one semester, the order of students' scores can differ from the absolute order. So we can use these data to simulate a contest, in which each semester acts as a judge assigning a score.

We use these data in our computer program and get the results of **Table 6**.

Table 6.
Results of analysis of departmental data ($P = 50$, $J = 3$).

W	T	papers selected
2	19	$P_{(1)}, P_{(3)}$
3	22	$P_{(1)}, P_{(2)}, P_{(3)}$

Generalization

How to Budget?

Funding for the contest constrains both the number of judges that can be obtained and the amount of time that they can judge.

Assume that each judge can mark n papers / day and the judge's salary is \$$s$ / day. We can hypothesize that funding f is a function of T, n, and J: $f = f(T, n, J)$, where T is total judging time. Obviously, $\partial f/\partial J > 0$, $\partial f/\partial T > 0$, $\partial f/\partial n < 0$, and $T = T(J)$. Fortunately, we have the an empirical formula

$$\left\lceil \frac{P}{J} \right\rceil + W + 2.$$

A reasonable functional form for f is

$$f = \left\lceil \frac{T}{n} \right\rceil \cdot J \cdot s = J \cdot s \cdot \left\lceil \frac{\lceil \frac{P}{J} \rceil + W + 2}{n} \right\rceil.$$

Since $k \le \lceil k \rceil \le k + 1$, we get

$$\frac{(P + (W + 2) \cdot J) \cdot s}{n} \le f \le \frac{(P + (W + 3 + n) \cdot J) \cdot s}{n},$$

which allows us to budget for the contest if the number of the judges has been given (see **Table 7**).

Table 7.

Cost of the contest, for various combinations of papers per day per judge and number of judges.

n	J	$\min f$	$\max f$
15	8	\$3,080	\$6,067
20	8	\$2,310	\$5,250
15	7	\$2,987	\$5,600
20	7	\$2,240	\$4,812

On the other hand, we can turn the equation around into the form

$$\frac{f \cdot \frac{n}{s} - P}{W + 3 + n} \le J \le \frac{f \cdot \frac{n}{s} - P}{W + 2}.$$

According to this, if funding is known, we can decide the number of judges (see **Table 8**). Of course, these are rough estimates.

Table 8.

Number of judges that can be hired, for various combinations of number of papers per day per judge and budget.

$P = 100$, $s = 350$, $W = 3$

n	f	$\min J$	$\max J$
15	\$5,000	6	22
10		3	8
15	\$7,000	10	40
10		7	20

Applying the Model to Different Kinds of Competition

For contests that give awards to just a few winners, our model is an effective and rational scheme. For contests that give various awards at different levels, we can modify a few parameters in our model. There are two methods.

- Method 1: Suppose that the contest committee expects to classify the participants into different levels in some given proportions, say four levels of 5%, 10%, 35%, and 50%, similar to the MCM. In the first round, we reject 50% as Successful Participation; reject 35% in the 2nd round as Honorable Mention; reject 10% in the third round as Meritorious; the remaining 5% are Outstanding.
- Method 2: We set the value of LEVEL as needed in each round to distinguish participants of different levels. This method is more flexible and fairer than Method 1.

Final Scheme

We summarize our final scheme:

- Divide the judging process into several screening rounds and follow the principles below in each round until W papers remain.
- Use a scoring scheme.
- Do not compare the scores from different judges.
- In the first round, distribute papers to all judges evenly. After scoring, select out the top $2W$ papers in each group to enter the next round.
- At the end of each round, for each paper, calculate the number of papers better than it (which we call the current rank of the paper), then reject every paper whose current rank is more than $W - 1$.
- At the beginning of each round, dispatch papers with a close current rank to the same judge, if possible.
- The number of papers distributed to each judge in each round should be as equal as possible.

Our Suggestions

- Properly reducing the number of rejected papers in the first round would decrease the error probability.
- Altering the number rejected in each round as needed is helpful in competitions that determine different levels of the participants.
- To be more practical and efficient, we suggest prejudging the papers at first, that is, rejecting the papers of distinctly poor quality.

- Between rounds, have some discussion among judges so that they gain some knowledge of the levels of papers as a whole. Such a feedback mechanism surely helps reduce the standard deviation of judgment.
- When there are about $2W$ papers left, all the judges gather to read the remaining papers together, if time permits, to select the top W papers.

Strengths and Weaknesses

Strengths

- We have shown how our model provides an efficient scheme in correct selecting winners. The model was not only tested in a computer simulation but also proved adaptable to real cases.
- Our model is also very stable. All parameters, which we set arbitrarily, can be changed without changing the quality of the model.
- We obtain from our model an empirical formula for T, based on P, J, and W.
- We use Bayesian estimation to take into account the differences among judges.
- Our model is flexible enough to be applied to different kinds of competitions.

Weaknesses

- We are unable to demonstrate that our model is optimal.
- We would like to be able to improve our estimation of the parameters for each judge.

References

Box, G.E.P. and Tiao, G.C. 1973. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Wang, Yihe. 1986. *Introduction to Discrete Mathematics*. Harbin, China: Harbin Institute of Technology Press.