Modeling Better Modeling Judges

Brian E. Ellis
Chad Hall
Charles A. Ross
Gettysburg College
Gettysburg, PA 17325

Advisor: James P. Fink

Summary

We designed a system for judging a contest of papers with two main goals in mind: to minimize the number of reads by a judge and to ensure a fair contest. We first designed a model that would best predict the choices of human judges comparing just two papers. The basic premise is that the closer two papers are in an absolute ordering, the more likely the order of the papers is to be reversed by a judge, whereas the farther they are apart, the less likely a reversal.

Our model accommodates arbitrary numbers of judges, papers, and winners. The P papers are split into S stacks. To ensure fairness, two judges read each stack. From every pair of stacks, W papers advance to the next round. If two judges cannot agree which W should advance, the Head Judge decides. The rounds continue until $2W$ papers remain, when a balloting process among four judges and the Head Judge determines the W winners.

We can predict the total number of reads made in the judging process and the maximum number of reads by any judge. We calculate an optimal number of judges so that all judges have nearly the same number of reads.

Testing on a computer, we found that our model fails to pick W out of the top $2W$ no more than 0.1% of the time. These failures are attributable to the human factor in judging. For the given problem of 8 judges deciding on 3 winners from 100 papers, our model predicts and tested at 254 total reads, with 32 papers read by each judge; the model fails to select 3 of the top 6 only 0.08% of the time. For 32 judges deciding on 7 winners from 350 papers, our model predicts and tested at 1,162 total reads, and 36 papers read by each judge; the model fails to select 7 of the top 14 only 0.01% of the time.

Assumptions with Justifications

- Papers:
  - Ranking: There is an absolute ordering of the papers, so we can determine if the winning papers are within in the top $2W$ total papers.

  - Number: The number of papers is far greater than the number of winners.

- Judges:
  - Knowledge: All judges are knowledgeable about the question posed and can easily determine if a paper has merit; otherwise, a paper cannot be fairly evaluated.
  - Preferences: All judges will agree on the ranking of a particular paper within some margin of error. Each judge has personal preferences about what is desirable in a paper. Also, when a judge is asked to read a large number of papers, there must be some margin of error in the ranking process.
  - Ability: A judge can read up to 20 papers at a sitting and still pick out the top papers with a reasonable amount of accuracy. In speaking with a number of professors and contest evaluators, we found that 20 is the upper bound on the papers that the professors feel that they can evaluate fairly at one time.
  - Head Judge: The Head Judge only settles disputes and votes in the final round; the Head Judge is not counted in the number J of judges.
  - Number: The minimum number of judges is 5, including the Head Judge. There must be enough judges to evaluate all of the papers fairly; the more judges, the more accurate the process will be.

- *Fairness* is the ultimate variable. In any contest, judges must be willing to sacrifice time and energy to ensure that the best papers will win the contest. The credibility of the contest is based upon the fairness and correctness of the judging.

Definitions of Constants and Terms

P: total number of papers

J: total number of judges, not including the Head Judge

J_k: representation of judge k

W: total number of winners

read: one judge reading one paper one time

round: a process of elimination in which a set of papers is cut to W papers

R_a: the representation of round a

S_a: the number of stacks in round a. A *stack* is a set of papers of size $< P$.

N: the number of papers in a stack

S_{jk}: representation of stack j in round k

error: a judge's ordering that contradicts the absolute ordering

The Paper Contest Model

The model begins by dividing the P papers into S stacks. Judges then perform an elimination round in which two judges work together to combine two stacks into one stack of W papers. The comparisons are made by rank ordering, using no numerical scoring system. The process is repeated for the new stacks until two stacks are left. The final round then enacts a voting process on the last two stacks to declare the winners.

Preliminaries

We first determine the number of stacks, S_1, needed for the first round. To ensure a symmetric elimination, we need S_1 to be a power of 2. By our assumptions, each judge can read up to 20 papers, so the size of a stack cannot exceed 20. The number of papers in each stack is $N = P/2^n$, where n is the smallest value that satisfies

$$N = \frac{P}{2^n} \leq 20.$$

If 2^n does not divide P evenly, N is rounded up. The papers are distributed as evenly as possible about the S_1 stacks. We assign each judge one stack until we run out of either stacks or judges. If we run out of judges, then some judges will be asked to repeat the first round.

First Round

Judges J_1 and J_2 are assigned stacks S_{11} and S_{21}. Judge J_i chooses W papers from the stack S_{i1}; W, to ensure that all of the W best papers cannot be eliminated in round R_1. Once done, they swap stacks. Judge J_1 then chooses W from S_{21}, while J_2 chooses W from S_{11}. Together, they compare their lists and determine W from the union of S_{11} and S_{21}. If there is a dispute, the Head Judge determines which paper advances. Each pair of stacks is cut to W papers in the same manner. At the completion of the first round, there are $S_2 = 2^{n-1}$ stacks and $N = W$ papers.

Why Choose W Every Time?

The scenario could arise in which the top $2W$ papers fall into one stack in any round. If we return any fewer than W papers, the model would automatically

fail. To return more than W papers would increase the stability of the model, but not to a degree that would warrant the increased number of reads required.

Second and Subsequent Rounds

There will be $n-2$ "middle" rounds (see **Appendix A**). The procedure for these rounds can be generalized with the introduction of a variable r that holds the value of the round number. At the beginning of R_r, we have $S_r = 2n-r+1$ stacks and $N = W$ papers. The next two available judges are assigned the stacks S_{1r} and S_{2r}. Each chooses W papers from the union of stacks S_{1r} and S_{2r}, and then they agree upon a final W to advance, with the Head Judge settling any disputes. Every pair of stacks is cut to W papers in the same manner. This is repeated round by round up to and including round R_{n-1}, at the completion of which there will be $2W$ papers remaining.

Final Round

The final round, R_n, is a voting process. To ensure fairness and to account for the importance of the final decision, we choose five judges, including the Head Judge, to evaluate the papers. These judges read the remaining $2W$ papers and rank-order them. An official, possibly an extra judge, tallies the votes, giving W points to first place, $W-1$ points to second place, and so forth, down to 1 point for place W. The W papers receiving the most points are the winners. If there are any ties in the points, the ballot of the Head Judge breaks the tie.

Human Factor

The one variable that this or any model cannot control is the human factor. We simulate this human factor by using a probability distribution that models what an actual judge may do. If all the judges are exactly the same, paper 1 will always be ranked ahead of paper 2. However, individual judges will have preferences about what they would like to see in a paper. The most common example is one judge who would weight presentation over substance while another judge would rate substance over presentation. In this case, paper 2 could easily be rated above paper 1. To model this factor, we chose the following function as the probability that a judge's ranking of two papers differs from the absolute ordering

$$E(P,d) = \frac{1.46 + \arctan(1 - 60d/P)}{2.92 + \frac{\pi}{2}},$$

where there are P papers in the contest and d is the distance between the two compared papers on the absolute scale.

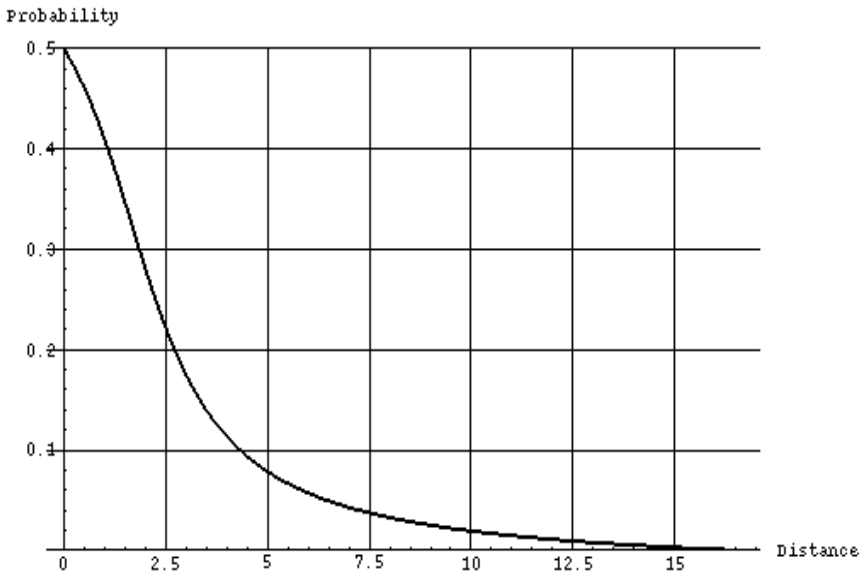

Figure 1. The operating characteristic curve for a judge's ranking of two papers. (Note that this is not a probability density function.)

The formula gives the probability of the judge making an error as a function of the true difference in ranks between two papers. As the distance between two ranks increases, the probability of reversing of the papers decreases quickly. The probability of error when there is a $0.01P$ difference between two papers is approximately 50%. So, the choice between papers 5 and 6 for $P = 100$ is completely random. The probability of error when there is more than a $0.17P$ difference is 0. In this case, the magnitude of the difference between the papers is too great; it would be impossible to err in the comparison. The values for the probability of error between $.01P$ and $.17P$ are representative of real life—the closer two papers are, the more likely a judge's personal preferences of style will influence the ordering of the papers. Similarly, the further two papers are apart, the less likely the judges preferences will be able to affect their comparison.

Results

Total Reads

The total number of reads, excluding arbitrations by the Head Judge, is given by

$$2P + \sum_{i=2}^{n-1} 2^{n-i+2}w + 5(2w).$$

The first term is to the number of reads in R_1, the second takes care of rounds R_2 through R_{n-1}, and the third is for R_n (see **Appendix A**).

Number of Judges

The model requires five judges, including the Head Judge. The model can accommodate any $J \geq 4$, but there is an optimal number of judges J_O that minimizes the maximum number of reads per judge. That optimal number of judges equals S_1, not counting the Head Judge. All J_O judges are needed in R_1, with one-half used in R_2, one-fourth in R_3, and so on. We use each judge in the first and in one subsequent round, leading to nearly the same number of reads (see **Appendix A**).

Maximum Reads per Judge

If $J \geq J_O$, the maximum number of reads is

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

If $J < J_O$, the maximum number of reads can be very large, even unreasonably large. In this situation, some of the J judges will be required to read two or more pairs of stacks in round one. Already these judges will have to read at least 40 papers and possibly more, because the second and subsequent rounds haven't even started. If a $J < J_O$ is chosen, J must be close to J_O or there will be many unhappy judges.

Testing the Model

We implemented the model in the C programming language, making some minor additional assumptions (see **Appendix B**).

We then ran tests for various combinations of P and W, always using the optimal number J_O of judges. We did 10,000 iterations each for the cases in **Table 1**. The test data returned an average failure rate of 0.023%.

Table 1.
Combinations of number of papers P and number of winning papers W, for each of which 10,000 iterations were run.

P	W	percentage failure	total reads	max reads per judge
50	2	0.02%	104	26
100	3	0.08%	254	32
200	5	0.01%	570	36
350	7	0.01%	1,162	36

The model is valid. It agrees with the formula for the maximum number of reads, total reads, and, most important, the final W are consistently among the top $2W$. The small failure rate is attributable to the human factor. Whenever the human element is involved, there are bound to be rare cases that occur.

Strengths and Weakness of the Model

Strengths

- The probability of the model failing is extremely low, usually less than 0.1%.
- The model takes into account the possibility of human error.
- All judging is done via direct comparison, and at least two judges must concur for a paper to advance. There is no numerical scoring, which can be biased by the grading scale of the judge; and an error by a judge has less of a chance of advancing an unworthy paper.
- Our model performs extremely well for the example posed in the original question ($P = 100$, $W = 3$, and $J = 8$) (see **Figure 2**). It fails only 0.08% of the time, while limiting the judges to 32 reads each (one-third of the total number of papers) and the total reads to 254.
- Most important, we would feel very comfortable using the model to judge our paper in the 1996 MCM. The model is fair. The top papers win virtually every time.

Weaknesses

- The model has definite bounds of effectiveness. We have set a bound on the number of papers that a judge can read at one sitting at 20. After the first round, judges read $2W$ papers each round. Thus, the number of winners must be less than or equal to 10. Allowing 2% of all papers to be winners, the total number of papers must be less than or equal to 500. A possible solution for large P is to break the contest into two halves of less than 500 each and run the model for each half.
- We had to model the human factor, which is difficult. The data on which we based the curve were our best estimate of human nature. We did not have data to consult to see how humans would actually perform under these circumstances. All of our testing and the validity of our results are based upon the assumption that our equation is actually representative of what occurs in the real world. If further research deems that equation to be inaccurate, it will be easy to adjust our model to a new equation.

Appendix A: Rationale and Proofs

There Are $n - 2$ Middle Rounds

In each round, exactly $W(S_r+1)$ papers advance into the next round, where $S_r + 1 = 2^{n-r}$. This is true because the number of stacks is halved with each

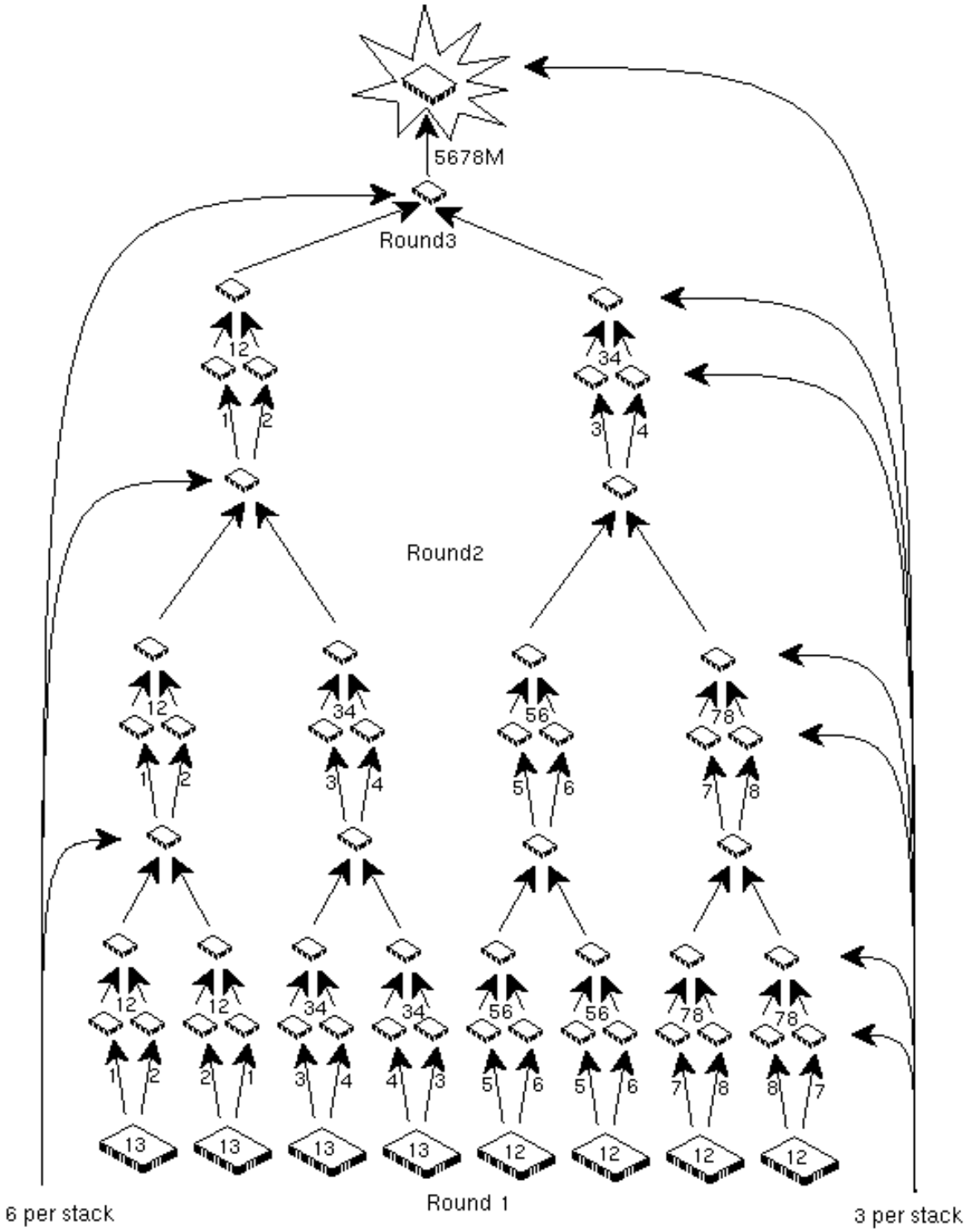

Figure 2. Diagram of the operation of the model for the original setting of $P = 100$, $W = 3$, and $J = 8$.

successive round, and there are W papers in each new stack. The final round begins when only $2W$ papers remain. We are in the final round when $2^{n-r} = 1$, or $2^n = 2^r$. So $n = r$ and the total number of rounds is n, including the first round and the final round. Hence there are $n - 2$ middle rounds.

Total Number of Reads

In R_1, every paper is read twice, yielding $2P$ reads. In the middle rounds, there are 2^{n-r+1} stacks of W papers each. Each stack is read twice, yielding $2^{n-r+2}W$ reads per round, for $n - 2$ rounds. Round R_n has 5 judges, each reading the final $2W$ papers. Thus, the sum of the reads for all rounds is

$$2P + \sum_{i=2}^{n-1} 2^{n-i+2}w + 5(2w).$$

Rationale for Optimal Number of Judges, J_O

We would like each judge to read approximately the same number of papers. This happens in the first round (some judges may read one additional paper if the stacks cannot be divided evenly). In each successive round, the number of papers read by a judge is $2W$. If the number of judges is $2n$, each judge is guaranteed exactly 2 rounds of judging. The number of judges needed is halved with each subsequent round. Thus, there will always be 4 judges for R_{n-1}, leaving 4 judges who have yet to judge a second time. These four judges plus the Head Judge make up the 5 who judge the final round. If J_O judges are used, each judge reads nearly the same number of papers, since every judge reads in the first round and one subsequent round.

Maximum Number of Reads for a Judge

If $J = J_O$, each judge reads in exactly 2 rounds: the first round, with approximately

$$2\left\lceil \frac{P}{2^n} \right\rceil$$

papers, and either a middle or a final round, with $2W$ papers. Thus, the maximum number of reads is

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

If $J > J_O$, some judges read in only one round. So long as $J < 2J_O$, at least one judge will read in two rounds, so the maximum number of reads is still be the same as above.

If $J < J_O$, some judges read more than one pair of stacks in the first round, possibly many pairs; the maximum number of reads could become very large.

Appendix B: The Computer Test Model

In using a computer to test a model, we must make some assumptions about human behavior that can be implemented by the computer. For the most part, our assumptions about human behavior are taken care of in the equation for the error factor. The other assumptions that we make are:

- Judges always require a third party to settle any disputes. It is very difficult to implement the power persuasion would have in a discussion between two judges.
- Judges are reasonable. If a judge has read two papers earlier and reads them again, they will again receive the same relative ranking, except possibly in the last round, which allows for more scrutiny to each paper.
- There is no Head Judge in the program. The person who settles any disputes is simply the next available judge who has never before compared the papers.
- The optimal number of judges is always used.