Practitioner's Commentary: Computer Support for the MCM

Steve Harper
Mathematics Dept.
Caroll College
Helena, MT 59625
sharper@saints.carroll.edu

Judging the math modeling contest is mostly a human endeavor, with a proper place for a computer to help. Since human judges are, by definition, human, there need to be ways to watch for human bias in order to get fair contest results.

The current scheme that Contest Director Frank Giordano uses has evolved to address these concerns. To make a fair judgment, each judge needs to see excellent, good, and average papers. There are too many papers for each judge to read them all, and assigning papers by random draw will not always give a good balance. After being ranked using the scores from prior rounds, the papers are distributed to ensure that each judge gets a variety in quality.

Since different people do not give the same score to the same paper, the scores are weighted (to account for natural human tendency to "grade low" or "grade high"). A more subtle problem is that people tend to "root" for a paper that may, for instance, resonate with how the judge would approach the problem (the "right answer," so to speak). Making the scoring scale 40–100, rather than 1–100, can reduce the impact of rooters who give low scores to other papers. Having the judge rank all the papers for that round also gives the contest director more control to ensure that a paper is not arbitrarily eliminated. For instance, if a paper with a low score has a ranking of 2nd out of 15 papers, it may deserve to stay for another round and another opinion. Furthermore, the ranking provides a way for judges to give a numerical score based on the established criteria yet still note that there are unquantifiable factors that make one paper rank better than another paper with a higher score.

The judging for this contest previously used a spreadsheet. This gave some last minute flexibility, at the cost of data entry errors plus late-round bleary eyes trying to line up too much numerical data that would not fit on one page. In 1996, the contest used a custom FoxPro database program with three database tables (for scores, for judge information and weightings, and for round names and elimination scores).

Judges are assigned using letter codes rather than numbers since numerical scores abound on the printed score sheets. Each judge needs a random set of papers containing both good and average papers (that this judge has not read before). After the two triage rounds, the remaining contestant papers are split into 4 stratified layers based upon accumulated weighted scores. The

computer tries to give an equal number of papers to each level and assigns papers so that each judge will get a variety of quality in papers. It also ensures that no judge reads the same paper twice. The contest director resolves any problems generated (such as the computer not assigning any paper to a judge if, by luck, the only papers left in the contest have all been read by that judge).

The computer prints out the judge assignments for the contest director in score order, and for the individual judges in document number order (so the judge doesn't know how the paper has fared so far).

The computer checks scores as they are recorded and alerts the human input operator for typing or recording errors. (The score that a judge intends to put on a paper should not be rendered invalid by a typo or two.) It notes problems if the human tries to enter a score for a document that doesn't exist or is already out of the contest. It verifies that the actual judge is the same as was assigned (though this can be overridden). Other problems noted are scores out of range, changes to an existing score ("Is this really a change, or did you type the wrong document number?"), and ranking within a round (saying "This is the 5th best paper out of 4" will not pass inspection).

Some individual judges tend to grade harder or easier than others. A weighting formula (from Frank Giordano, who inherited it from Ben Fusaro) is designed to try to account for that with a minimum of effort. The formula is:

$$\text{Weighted Score} = \begin{cases} \dfrac{\text{Population Mean}}{\text{Judge's Mean}} \times \text{Raw Score}, & \text{if Population Mean} \le \text{Judge's Mean;} \\ \dfrac{100 - \text{Population Mean}}{100 - \text{Judge's Mean}} \times (\text{Raw Score} - 100) + 100, & \text{otherwise.} \end{cases}$$

(Note: no weighted score can be greater than 100, no matter how tough the judge is on the other papers.)

To allow for human modification, this weighting is applied in two steps. The program first calculates the judge's ratio, and the contest director can then individually adjust the judge's weighting. For example, a judge could read only a couple of papers before getting an emergency phone call and having to leave, and the weighting may be too far off (in the opinion of the contest director) to be useful.

Next the weighting ratio that goes with each judge is applied to each document that the judge scored. Again, the contest director can adjust the weighted score for any one paper, if there is good reason to do so.

To calculate the weighting, the only papers considered are the ones still left in the contest. All prior rounds for those papers are considered in calculating the weighting for each judge. Then the new weighted score is figured retroactively for the prior rounds to obtain a new weighted total. Thus, in each round, the weighting has to be recalculated. (Note: In each round, the low-scoring papers drop out. While the number of recorded scores is increasing, the number of papers left is decreasing, so the total number of scores included may go up or down. There always is a concern with weighting too small a sample.)

The length of time that a judge spends on the paper is also a factor, since a score for a Triage Round reading of 5 minutes does not deserve the same confidence as a 30-minute reading in the Finals.

Then, based on the weighted score, the computer suggests which papers to eliminate. (The elimination scheme is stored in the database and is easy to change.) It calculates what score will leave a (preselected) percentage of the original contestants after the current round, compares that to a (preselected) minimum score, and suggests the higher number.

After review, the contest director can decide to draw the elimination line in a different place, as well as check the set of scores for individual papers below the line to decide which papers deserve another chance. (If a paper got a 99 and a 40, it would probably deserve another read before being tossed out of the contest.)

The percentage and minimum scores for the 1996 contest are shown in **Table 1**.

Table 1.
Percentage and minimum scores for judging rounds of the 1996 contest.

Round	Score range	Minimum continuation score	Percentage remaining
Triage 1	0–7	1	99%
Triage 2	0–7	6	90%
Screening	0–7	12	43%
Final 1	20–50	25	30%
Final 2	40–100	85	18%
Final 3	40–100	150	10%
Final 4	40-100	220	6%

The computer also records the final classification based on when the document is eliminated:

- Survive Final 4: Outstanding
- Survive Final 3: Meritorious
- Survive Final 2: Meritorious
- Survive Final 1: Honorable Mention
- Survive Screening: Honorable Mention
- Survive Triage 1: Successful

The contest director can review these classifications and change them.

To help the contest director make elimination decisions, the computer prints out reports in weighted score order. At all times, a printout is available (in document number order) to answer the question, "Whatever happened to paper number such and such?"

Using this database program solves many problems. However, as one would expect, there is a cost—flexibility. The data are not readily accessible to noncomputer folks, so changing the structure of how the contest is judged is not too easy. It was not feasible in the time allotted to make a program generally usable for every judging situation possible. Given the stability of the contest over the last several years plus time constraints, trading future flexibility to solve present needs (for data entry, judge assignment, score weighting and elimination) seemed a fair trade.

Are there improvements? Are you kidding? The final round ends with judges arguing over the top 3% of papers about which ones are really Outstanding and which of those deserve special awards. For all judges to participate, each needs to have read several of the top papers—more than just luck would allocate. Next year's program already has a better judge assignment algorithm in place!

About the Author

Steve Harper teaches at Carroll College in Helena, MT, where students are encouraged to have a variety of background work in areas other than just computing. (Steve himself has a background in accounting, politics, wind energy, and consulting.) He designed and coded the database and program for the 1996 contest.