Rating Skating

Gilbert W. Bassett, Jr.; Joseph Persky

Journal of the American Statistical Association, Volume 89, Issue 427 (Sep., 1994), 1075-1079.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199409%2989%3A427%3C1075%3ARS%3E2.0.CO%3B2-W

Rating Skating

Gilbert W. BASSETT,* Jr. and JOSEPH PERSKY*

Among judged sports, figure skating uses a unique method of median ranks for determining placement. This system responds positively to increased marks by each judge and follows majority rule when a majority of judges agree on a skater's rank. It is demonstrated that this is the only aggregation system possessing these two properties. Median ranks provide strong safeguards against manipulation by a minority of judges. These positive features do not require the sacrifice of efficiency in controlling measurement error. In a Monte Carlo study, the median rank system consistently outperforms alternatives when judges' marks are significantly skewed toward an upper limit.

KEY WORDS: Breakdown; Majority rule; Median; Ranks.

1. INTRODUCTION

Early during the 1992 Winter Olympics, Scott Hamilton, the former Olympic champion now working as an announcer for CBS, made a valiant but unsuccessful try to explain how judges' marks are aggregated to determine the placements of figure skaters. Once burned, he gingerly avoided the subject for the duration. Hamilton's difficulties reflected the complex and at first glance arcane procedures dictated by the rulebooks. Yet officials of the United States Figure Skating Association (USFSA) and the International Skating Union (ISU) have long claimed that their approach to judging avoids arbitrary mechanisms in favor of the logic of majority rule. The purpose of this article is to explore this claim.

Figure skating is one of the most graceful and aesthetic of sports. It also involves quick movements and subtle variations. These characteristics of skating, similar in many respects to diving and gymnastics, make the appropriate ranking of competitors quite difficult. Not surprisingly, all three of these sports rely on expert judging to determine placements in competitions. Unwilling to trust such responsibility to a single individual, the rules call for a panel of judges. Yet using such a panel creates a difficult and interesting question: how best to combine the judges' marks?

In considering this question, it seems reasonable to first ask why judges' rankings should ever differ. Why don't all the judges in a competition agree? What we want is a "model" of the generating process that gives rise to such differences. Unfortunately, even a cursory acquaintance with figure skating suggests a whole spectrum of possibilities.

At one extreme, we might imagine that the quality of each skater's performance is an objective entity subject only to errors in measurement. In such a view, judges' marks differ in the same way as clocks timing a swimming event might differ. This judge might have blinked at an important moment, or that judge might have concentrated on one member of a pair of skaters just when the partner wobbled. In this model, judges are imperfect measuring devices. The aggregation problem here is essentially one of measurement error.

At the other extreme, the differences among judges might represent not errors of measurement but rather genuine differences in tastes. Judges' preferences would presumably reflect real aesthetic differences, although they might also be influenced by national pride and other less relevant motivations. In a world of complex aesthetics, we face all the aggregation problems of collective decision making (see Arrow 1963 or Sen 1970). Moreover, skating officials worry continually about the strategic behavior of their judges. Where judges hold strongly to their preferences or have a personal stake in the outcome, we must also worry about the distortions that can be produced by strategic voting behavior.

Both of these models are obviously oversimplifications. Yet even at this level, it is easy to appreciate that a system of aggregation that showed highly desirable properties with respect to one of these might fail to perform well with respect to the other. But both models have some plausibility; at any given competition, one or the other might dominate. At prestigious competitions such as the Olympics or the World Championships, judges' marks are very likely influenced by both tastes and problems of measurement. At local or regional meets measurement problems probably dominate. Any system of aggregating judges' rankings must be considered in light of both models. Thus the rating skating problem requires us to search for a set of aggregation criteria relevant to both measurement error and preference aggregation.

Skating officials have long maintained that their placement system was desirable because it embodied the principle of majority rule. Although the concept of majority rule is open to a number of interpretations, we show that the system that has been adopted is essentially a median rank method with tie-breaking rules. The identification of the method with median ranks does not seem to have been noticed in the skating literature. Further, although the system has evolved only informally, we show that it is actually the only method that satisfies a majority rule and incentive compatibility (or monotonicity) requirement in which a skater's final rank cannot be decreased by a judge who gives the skater a better mark. Hence the skating associations have settled on the unique aggregation method of median ranks, which is resistent to manipulation by a minority subset of judges and also satisfies a reasonable incentive compatibility requirement.

These results relate to the problem of aggregating tastes. One might expect that a ranking system well tuned to handling such aggregation might perform poorly when evaluated

* Gilbert Bassett and Joseph Persky are Professors of Economics, University of Illinois at Chicago, Chicago, IL 60607. The authors would like to thank two anonymous reviewers, Benjamin Wright of the International Skating Union, Dale Mitch of the United States Figure Skating Association, Tammy Findlay, Shiela Lonergan, Nicole Persky, and all the mothers of the McFetridge Ice Skating Rink in Chicago.

Journal of the American Statistical Association
September 1994, Vol. 89, No. 427, Statistics in Sports

Skaters	Judges 1	2	3	4	5	6	7	8	9	Place of Lowest Majority	Size of Lowest Majority	Final Place
A	5	2	2	2	4	4	1	2	4	2	5	3
B	2	1	3	1	2	2	3	1	3	2	6	2
C	3	3	5	5	3	3	2	5	2	3	6	4
D	1	6	1	3	1	1	5	3	1	1	5	1
E	6	5	4	4	6	6	4	4	6	5	5	5
F	4	4	6	6	5	5	6	6	5	5	5	6

Figure 1. Hypothetical Ordinals for a Component Event.

from the perspective of measurement error. Yet we find that the official aggregation procedures deal effectively with the persistent measurement problem of "score inflation" that skews marks toward the upper limit of the scoring range. Thus we conclude that even in competitions where no manipulation is likely but judge's marks are subject to random error, the system performs well.

In Section 2 we provide a brief overview of the official scoring system used by the ISU and the USFSA. We explain that the system is essentially median ranks, but with somewhat arbitrary rules for breaking ties. In Section 3 we show that the median rank method can be justified in terms of a majority rule incentive requirement. In Section 4 we go on to consider the relative performance of the system compared to alternatives in the context of a simple model of error generation. Finally, we summarize results in Section 5.

2. THE RULES

At all major USFSA events, as well as the World Championships and the Olympics, a skater's place is determined by a weighted average of two component events. The short, or original, program is weighted one-third, and the long free-skating program is weighted two-thirds. Each component event is scored by a panel consisting of nine judges. At lesser competitions, there are still three component events: compulsory figures (20%), original program (30%), and free skating (50%). Moreover, there often are fewer judges (but always an odd number) at such events. The compulsory figures component has been dropped from more prestigious competitions, because the slow and tedious etching of school figures makes a less than dramatic video scene.

For each component, a judge gives two cardinal marks on a scale of 1 to 6. For the original program the marks are for required elements and presentation; for free-skating the marks are for technical merit and composition style. These marks are the ones displayed prominently at competitions. But there is a long trail from marks to placement.

Take, for example, the placements in the original program. (Placements in the free-skating program are determined in exactly the same way.) First, for each judge an ordinal ranking of skaters is determined from total marks, the sum of the two subcomponent cardinal scores. These ordinal ranks and not the raw scores become the basis for determining placements.

As presented by the USFSA rulebook, the procedure continues as follows: "The competitor(s) placed first by the absolute majority (M) of judges is first; the competitor(s) placed second or better by an absolute majority of judges is second, and so on" (USFSA CR 26:32). Note here the expression "second or better." In calculating majorities for second place, both first and second ranks are included. In calculating a majority for third place, first, seconds, and thirds are included, and so on for lower places. If for any place there is no majority, then that place goes to the skater with a majority for the nearest following place.

Now of course below first place, there can be numerous ties in this system. (If a judge has given more than one first because of a tied mark, then there can also be a tie for first.) The basic rule for breaking ties is that the place goes to the competitor with the greater majority. If after the application of the greater majority rule there is still a tie, then the place goes to the skater with the "lowest total of ordinals from those judges forming the majority." And if this does not work, then the place goes to the skater with the lowest total of ordinals from all judges. In all cases of ties, the skaters involved in the ties must be placed before other skaters are considered.

To demonstrate how this all works, consider Figure 1, which contains hypothetical ordinal rankings for a component event. Notice that the tie for second between A and B goes to skater B because of the greater size of B's majority for second. Skater A then gets third place, because A must be placed before anyone else is considered. Because no one has a majority of fourths, we go on to consider E and F, each of whom has a majority of fifths. Because each has five judges in their majority, breaking the tie depends on the sum of ordinals in each majority. E then wins with the lower sum, 21, as compared to F's 23.

After just a bit of reflection, it is clear that the placement system used in figure skating starts from a ranking of the median ordinals received by skaters. As defined by the rules, a skater's initial placement depends on the "lowest majority." But this "lowest majority" is just equal to the median ordinal. A majority of judges ranked the skater at the skater's median ordinal or better. It is true of course that a number of tie-breaking devices are applied. These rules involve several other concepts. But under the current procedures, a skater with a lower (better) median will never be ranked worse than one with a higher (worse) median. Such a result is explicitly ruled out, because all tied skaters must be placed before any remaining skaters are considered. In particular, all skaters tied at a given "lowest majority" or median rank must be placed before any other skaters are considered. Notice that in the absence of this rule, a reversal vis-a-vis the median rank rule could easily occur. For example, referring to Figure 1, if after failing in a tie-breaking situation for second place, skater A had to compete with skater C for third place, then the winner would be skater C (despite A's median of 2) because of a greater majority of "3s or better"; C has six "3s or better," whereas A has only five.

Although over the years there have been a number of changes in the various tie-breaking mechanisms, since 1895 the ISU has used its concept of majority rule to determine placements. The only exception we have discovered was an experiment in 1950 that used a trimmed mean. The system has now evolved to a point where it is clearly one of median ranks.

3. MEDIAN RANKS AND MAJORITY RULE

Why use the median rather than the average ordinal, the sum of the raw scores, or a trimmed mean? As in other sports involving subjective judging, ice skating has been plagued by charges of strategic manipulation. This problem is a common one in the theory of constitution building. (There is of course a large literature addressing this issue; see Arrow 1963.) The most obvious reason for using medians is to limit the effect of one or two outliers on the final rankings. But there are any number of ways to begin to guard against such manipulation. In defense of their system, skating officials from the ISU and USFSA have often claimed that it embodies the essence of majority rule. The heart of their argument is that a skater ranked best by a majority of judges should be ranked best overall.

In addition to its relation to majority rule, a system of median ranks has at least one other attractive property: If an individual judge raises a skater's mark, then that action will never decrease that skater's placement. Thus if a judge raises the mark of one skater, that skater will either move up or stay the same in overall placement.

These two properties are attractive characteristics of median ranks that suggest it for serious consideration. But in fact we can make a stronger statement. If these two simple conditions are considered to be necessary, median ranking is the only system that will satisfy both. The result follows from the median as a high-breakdown estimator and the fact that such estimates satisfy an exact-fit property that is equivalent to a majority requirement in the aggregation context (see Bassett 1991).

To formally demonstrate the result, let $m_j(s)$ denote the raw mark and let $r_j(s)$ denote the rank of the sth skater by the jth judge, where $s = 1, \ldots, S$ and $i = 1, \ldots, J$. We suppose that higher-valued cardinal marks are assigned to better performances, and skaters are then ranked with "1" as best. (We assume that there are no tied marks, so that the marks yield a complete ordering for each judge.) The final rank of the sth skater is denoted by RANK(s).

An initial ranking is determined by a place function, denoted by P. The P function takes the matrix of marks and produces a vector, $\mathbf{p}$ with elements $p(s)$, which provides a partial order of skaters. The ranking RANK(s) is obtained by breaking the ties of $\mathbf{p}$.

The total mark is a particularly simple example of a P function. Here $\mathbf{p}(s) = (m_1(s) + m_2(s) + \cdots + m_J(s))$. Observe that this rule can be "manipulated" by a single judge. The skater from the "good" country who is clearly best in the eyes of all but the judge from the "bad" country can lose a competition if the "bad" judge gives that skater a very low mark. A trimmed mean is also a placement function, and of course trimming can eliminate the influence of a single "bad" country judge. But despite this, trimming can still violate our conception of majority rule.

We now formalize the requirements of a place function:

1. Incentive compatibility. A skater's final rank cannot be made worse by a judge who improves the skater's mark. In terms of P functions, this says that if $d > 0$ and $m_j(s) + d$ is substituted for $m_j(s)$, then $p(s)$ cannot fall.
2. Rank majority. If the rank matrix is such that skater s has rank r_0 for at least half the judges and skaters s' has rank q_0, for at least half the judges, where $r_0 < q_0$, then $p(s) < p(s')$.

Note that the rank majority requirement considers only situations in which more than half of the judges agree on the precise rank of skater s and more than half (not necessarily the same "more than half") agree on the precise rank of skater s'. The rank majority sets no explicit conditions on any other situation.

Many placement functions meet requirement 2; for example, the shortest half or least median of squares (LMS (see Rousseeuw 1984). The LMS identifies for each skater the half subset with the most similar or closest ranks and assigns as an initial placement function the midpoint of that interval. Clearly this satisfies the rank majority rule; but it does not satisfy Requirement 1. To illustrate this fact, consider a skater with the following ranks given by five judges: 1, 1, 3, 4, and 7. With LMS, this skater's placement function value is 2. But if the last judge improves the seventh place rank to a fourth place finish, then the skater's placement function actually falls to 3.5.

Theorem. Any place function that satisfies Requirements 1 and 2 is equivalent to the median rank place function.

Proof. It is easy to see that the median satisfies Requirements 1 and 2. To see that only the median rank and no other placement function satisfies these two requirements, we proceed by contradiction. Let M and R be marks and ranks evaluated by a P function satisfying Requirements 1 and 2, where

$$p(1) \leq p(2) \tag{1}$$

but

$$\text{med}_j\{r_1(1), \ldots, r_J(1)\} = x_0 > y_0 = \text{med}_j\{r_1(2), \ldots, r_J(2)\}. \tag{2}$$

We are going to change these marks without affecting either the relative placement of skaters 1 and 2 or their median ranks; however, after the change, a majority of judges will have given an identical rank score to skater 1 that is greater than an identical rank score given by a majority of judges to skater 2. But this will violate the majority requirement of a place function.

Consider the set of judges whose rank for skater 1 is $\geq x_0$; notice that this set includes a majority of judges. For each such judge, adjust marks so that (a) if $r_j(1) = x_0$, then do nothing; leave the mark and rank at their original values, or (b) if $r_j(1) > x_0$, then increase skater 1's mark so that the rank is decreased to x_0. It can be verified that this remarking and reranking leaves the median relation (2) unchanged, and, because the rank value for skater 1 goes down—the relation (1) also still holds (by the incentive requirement). Further there are now a majority of judges for whom the rank of 1 is x_0.

We now perform a similar operation for skater 2. Consider the set of judges whose rank for skater 2 is $\leq y_0$; notice that this set includes a majority of judges. For each judge in this majority set, (a) if $r_j(2) = y_0$, then do nothing, or (b) if $r_j(2) < y_0$, then decrease skater 2's mark so that her rank is decreased to y_0. It can again be verified that this does not change either (1) or (2). Further, there now is a majority of judges for whom the rank of skater 2 is y_0. Hence, by majority rule, $p(1) > p(2)$, which contradicts (1) and completes the proof.

We conclude that the median rank is the only placement function that possesses these two desirable properties. Of course, median ranks cannot perform miracles. Like all social welfare functions, this choice rule will, under specific circumstances, violate Arrow's list of properties. In particular, the winner of a competition as judged by USFSA rules can easily depend on "irrelevant alternatives." A new entrant into a competition can change the outcome, just as a spoiler entering a three-way election can upset a favored candidate.

At the same time, we should also note that our choice of Requirement 2 to represent majority rule is subject to dispute. This is only one of the possible interpretations of majority rule. Indeed the more familiar representation of this concept performs pairwise comparisons between alternatives. If the majority prefers **x** to **y**, then society prefers **x** to **y**. This is a different idea of majority rule than that contained in median ranks, and it is easy to construct examples (see Fig. 2) in which a majority of judges prefer **x** to **y** but **x** obtains a worse median rank than **y**. The well-known problem here is that such a ranking generally will not be transitive.

4. RANKING AS A MEASUREMENT ERROR PROBLEM

The median ranks used in placing figure skaters capture an interesting meaning of majority rule and offer obvious advantages in limiting strategic manipulation. Yet in the vast majority of competitions where there is little concern with such issues of preference, one can reasonably ask whether the present system is unnecessarily cumbersome or worse. For most competitions, the problem is not one of preference aggregation but rather one of statistical estimation, where concern is measurement error. Our first thought was that in these settings, the USFSA system would be less attractive than simpler aggregates, because its emphasis on median ranks ignores considerable information in determining placement. To look at this question, we conducted a series of Monte Carlo experiments comparing the official system to one of simple addition of cardinal marks. For completeness, we also included a trimmed mean similar to that used in diving competitions.

Skaters were assigned a "true" point score, which in turn defined a "true" ranking. The scores measured by individual judges were set equal to the true score plus a random error term. A normal error distributions was used, but as in actual meets, all scores were truncated at 6.0. As a simple measure of how well a system did, we calculated both the proportion of times that it picked the true winner and the average absolute error of placements.

In our first set of meet simulations, we treated the competition as consisting of only one component event judged on a simple six-point scale. Each meet consisted of five judges and six skaters (one through six), with true scores ranging in .2-point intervals from 5.8 to 4.8. The random error was taken to be normal with mean 0 and variance 1. (But as noted earlier, judges' scores were truncated at 6.0, thus skewing the distribution of scores). We ran 20,000 "meets" of this type. The simple addition of judges' scores correctly identified the true first place in 46% of the meets. But the USFSA system picked the correct first place finisher in 54% of meets. The trimmed mean did about the same as the sum, picking 45% of the correct first place finishers. The straightforward sum of ranks had an average absolute error in the estimated rank of a skater of 1.10, a figure identical to that for the USFSA system. The trimmed mean did only a tad worse, with an average absolute error of 1.12.

The result surprised us initially. But in hindsight, we realized that the success of the median ranking system was largely due to the mark ceiling imposed on the judges. In this situation, downward measurement errors for a good skater cannot easily be offset by upward measurement errors. Hence the average or total judges' score of a very good skater is systematically biased downward.

To demonstrate, we redid the simulation, but this time the highest skater had a true score of only 3.6 and the other skaters had scores again at .2-point intervals. The result, as we now expected, was that the USFSA system found a lower

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
.					
.					
.					
fourth place finisher	4	4	4	7	6
fifth place finisher	3	3	5	5	5

Figure 2. Conflicting Conceptions of Majority Rule. Notice that every judge but Judge 3 prefers the fifth place finisher, who has a majority of fives, to the fourth place finisher, who has a majority of fours. Also, in this case the fifth place finisher has a better (lower) total score and a better (lower) trimmed mean.

percentage of appropriate winners than the total system (46% vs. 48%). The trimmed mean also came in at 46%. The average absolute error in placement was now a good deal higher for USFSA, 1.15, as compared to 1.07 for the sum of marks and 1.11 for the trimmed mean.

Although hardly conclusive, these simulations suggest that the USFSA system may actually help in distinguishing among skaters of different performance levels when questions of preference are not seriously at issue. This result depends critically on the mark ceiling of six points, which strongly skews judges' marks. The median rank method works well with skewed scores.

5. SUMMARY

Like gymnastics and diving, figure skating requires a method to aggregate judges' marks. Unlike other judged sports, however, figure skating has adopted a system based on median ranks. Skating officials have often bragged that their system represents majority rule. We have shown that median ranks uniquely captures an important meaning of majority rule and provides strong protection against manipulation by a minority of judges.

One might have expected that these positive features would have required the scoring system to sacrifice efficiency in the more mundane world of measurement error. Yet, somewhat accidentally as the result of persistent mark inflation, we find that median ranks do a better job in controlling measurement error than two alternatives, total marks and the trimmed mean.

Although we can find no historical evidence that skating officials ever had this end in mind, they have picked a system particularly well suited to serve as both a method of statistical estimation and a means of preference aggregation as the situation warrants.

[Received April 1993. Revised November 1993.]

REFERENCES

Arrow, K. (1963). *Social Choice and Individual Values* (2nd ed.), New York: John Wiley.

Bassett. G. W. (1991). "Equivariant, Monotone, 50% Breakdown Estimators," *The American Statistician,* May, 135–137.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79,

Sen, A. (1970), *Collective Choice and Social Welfare,* Oakland: Holden-Day.

United States Figure Skating Association (1992), *USFSA Rulebook,* Colorado Springs, CO: Author.