EM0210: The Globe and Mail, February 16, 2002, page F6

New scoring system would have iced the gold for the Canadians

It's the eternal question when it comes to skating competitions: Can the bias of judges be detected and eliminated? **STEPHEN STRAUSS** finds that something called the Rasch measurement may do the trick

While much of the television viewing world gagged when Jamie Salé and David Pelletier were judged Olympic silver but not gold medal skaters this week, a small subset of statisticians was probably giving a silent yippee.

Another scoring controversy means another batch of numerical papers attempting yet again to resolve figure skating's eternal question: How do we deal with that rogue Russian judge?

This is not exaggeration: In the early 1990s, an entire session of an American Statistical Association's annual meeting was devoted to ice-skating judging.

The "Russian judge" in this construct is not literally a Russian, but simply any judge whose impartiality is overwhelmed by one or another non-skating factors. In the era of the Iron Curtain, it was widely assumed by Western skaters that Eastern European judges had colluded to fix the events for Russian or other Communist country skaters.

So the question becomes: Can one make a scoring system so refined that it catches vote-rigging and compensates for it? The answer appears to be both an unqualified "yes" and a qualified "maybe not."

The unqualified "yes" is as follows. Computerized statistical techniques that are specifically designed to uncover and compensate for judging prejudice already exist. And some have already been applied in a *post facto* sense to figure skating.

For instance, in 1997, Marilyn Looney, who is in the department of physical education at Northern Illinois University, took the highly disputed results of the 1994 women's Olympic skating final and subjected them to what is known as the Rasch measurement.

This statistical weighting formula, first proposed by Danish statistician Georg Rasch in the 1960s, has achieved great currency in educational testing, in which it is used to smooth out the effects of hard and easy graders on medical or dental certification exams.

What the program tries to do is break down the judging process into discrete elements that take into consideration the judges' idiosyncrasies and the program's technical difficulty. Maverick ratings can be flagged and their influence on the final scores calculated.

Looney reported that in a Rasch-adjusted universe second-place winner American Nancy Kerrigan should have beaten out Ucrainian Oksana Baiul for the gold medal.

What the program tries to do is break down the judging process into discrete elements that take into consideration the judges' idiosyncrasies and the program's technical difficulty. Maverick ratings can be flagged and their influence on the final scores calculated.

... the confidence you have in any adjustment is a function of the confidence you have that you have truly measured judge prejudice in the first place. Differences in educational judging severity might make use of hundreds of thousands of scoring patterns.

Four judgements – two each for the short and long programs – on 27 skaters may simply not be a large enough sample size for a computer to make readjustments free from its own statistical biases.

In answer to the question of what a Rasch-type analysis would have meant for the Salt Lake City Olympics, Mike Linacre, a U.S. statistician who was Looney's consultant for the earlier work, has done a simple computation of the "prejudice pattern" at work in the judging.

Taking into consideration the average score for each program, each task, each judge and each competitor, Linacre produced for *The Globe and Mail* a statistical picture of bias. Judges from four countries – Russia, France, Ukraine and Poland – were overwhelmingly lenient on Russian skaters and just as overwhelmingly tough on Canadian skaters.

"You can see there is a block vote by four judges and it is also surprising to see that the Polish and the Ukrainian judges were more lenient on the Russians than on their own skaters," Linacre says.

What the initial pass at the Rasch program does not tell you is why this pattern occurs.

In 1996, Montreal statisticians Bryan Campbell of Concordia University and John Galbraith of McGill reported that their investigation of Olympic figure skating stretching from 1976 to 1994 indicated that all judges tended to favour skaters from their own country, but that nationalism was inversely related to the ability of the skaters being judged.

If a skater was one of the top six in his or her class in the world, that excellence twigged national prejudice. However, if "you didn't have a shot at any of the medals, the judges appeared to be unbiased," Galbraith said.

In spite of all these caveats, the question remains: Could better number crunching have caught and corrected the prejudice pattern that Linacre turned up? According to at least some statisticians, if the Salt Lake City scores had been sent through a computerized Rasch prejudice filter, the newest Olympic figure-skating fiasco might never have occurred.

"The outcome could have been adjusted for judges' severity in real time after all the scores were in. We do it all the time in educational testing," says Mary Lunz of Measurement Research Associates, a Chicago company that uses the Rasch approach in medical certifications.

(*continued overleaf*)

The result would have been a process in which even the judges who judged could not have known for sure who had won until after their bias was corrected for.

And what would a program that adjusted for human error – either intended or accidental – have produced in medal rankings? "Thought you would ask this. Yes, a Canadian victory," says Linacre, who is working for a company that is developing Rasch software.

But maybe not, other statisticians say.

McGill's Galbraith points out that the confidence you have in any adjustment is a function of the confidence you have that you have truly measured judge prejudice in the first place. Differences in educational judging severity might make use of hundreds or thousands of scoring patterns.

Four judgments – two each for the short and long programs – on 27 skaters may simply not be a large enough sample size for a computer to make readjustments free from its own statistical biases.

All of this ignores the fact that massive cheating simply skews everything. "If all the countries put in the fix, I don't know of a model that can pick that up", Looney says.

Finally, while a more sophisticated statistical approach might make it harder to cheat, it may not resolve what has been put forward as the fundamental judging problem in skating. Scoring figure skating, the old saying goes, is like asking someone after seeing a ballet: Who won?

This is something that University of Chicago economist Gilbert Bassett has thought hard about. In 1994, in conjunction with colleague Joseph Persky, he did an analysis of the present International Skating Union scoring system. Their examination was inspired by what Persky thought was nonsensical statistical flaws in an ice-skating competition in which his daughter had participated.

"We initially thought, 'What a crazy system they have in place'. But after looking hard at it, we arrived at a completely different take on the matter," Bassett says.

They concluded the present system, in which, no matter how many low rankings he received, the skater who gets the majority of the first-place votes wins, is less prone to bias manipulation than scoring systems like the one used in gymnastics that knock out the top and bottom scores, or other ones that take into consideration total marks.

"Although we find no historical evidence that skating officials ever had this in mind, they have picked a system particularly well-suited to ..... serve as a method of statistical estimation", the U.S. pair wrote at the time.

As a fallout of their initial work, Bassett subsequently was part of a group of statisticians who had figure-skating judges explain to them what the judging judged. What emerged was a sort of culture of incomprehension. The statisticians easily understood how to interpret the numbers, but they didn't at all understand how the numbers were arrived at in the first place.

"Figure skating seems to me to be unjudgable, but these guys came in and talked as if they were Inuit and they had five different words for snow. As if there was objective material that anyone with reasonable abilities could say that was the guy who won, he says with amazement attaching to every word.

"To me, beauty is in the eye of the beholder."

1 Discuss critically the statement in the second paragraph of the right-hand column above: *... they have picked a system particularly well-suited to ..... serve as a method of statistical estimation,"* bearing in mind *our* two definitions:

* **Measuring:** the process used to determine the value of a variate;
* **Estimating:** a process which uses statistical theory to derive the distribution of an *estimator* and data to calculate an (interval) *estimate.*

The article EM0210 reprinted overleaf and above is used in Chapter 6 of the STAT 231 Course Materials and in Statistical Highlight #42, together with articles EM0303, EM9810 and EM0208.

2002-02-20