Judgment and Decision Making, Vol. 12, No. 4, July 2017, pp. 415–419

Order effects in the results of song contests: Evidence from the Eurovision and the New Wave

Evgeny A. Antipov* Elena B. Pokryshevskaya†

Abstract

The results of song contests offer a unique opportunity to analyze possible distortions arising from various biases in performance evaluations using observational data. In this study we investigate the influence of contestants' order of appearance on their ranking. We found that, in the New Wave Song Contest, expert judgments were significantly influenced by the contestant's running number, an exogenous factor that, being assigned randomly, clearly did not influence the output quality. We also found weaker statistical evidence of such an ordering effect in Eurovision Song Contest finals of 2009–2012. Keywords: ordering effects, cognitive bias, Eurovision, inter-rater agreement, judgment, ranking, voting

1 Introduction

Order effects can be a source of economic inefficiency in contexts where the quality of several candidates needs to be compared (Haan, Dijkstra & Dijkstra, 2005). Examples include job interviews and the grading of exams. In a health economics study it was found that the public's willingness to pay for three different health programs appeared to depend on the order in which they were presented to respondents (Stewart, O'Shea, Donaldson & Shackley, 2002): the first program in any sequence enjoys the highest willingness to pay. The authors give a possible explanation: respondents may feel that they have met their social obligations once they have contributed to the first program on the list.

Glejser and Heyndels (2001) and Haan et al. (2005) came to a conclusion that the order in which contestants perform in music competitions has a systematic influence on the final rankings, implying inefficiency in the jury's decision making process. Glejser and Heyndels (2001) used the data from the Elisabeth International Music Competition, while Haan et al. (2005) used data from the Eurovision Song Contest. In each of these competitions serial position was determined through a random draw, but contestants who performed later in the sequence generally received higher scores. Any effect of the order in which people are assessed on performance evaluation means that the evaluation process is biased (Page & Page, 2010). The results of the above-mentioned studies agree with psychological studies that tested order effects in sequentially judged options (Bruine de Bruin, 2005; Bruine de Bruin & Keren, 2003) and can possibly be explained with the help of Tversky's contrast theory (Tversky, 1977), according to which, when the subject focuses on a particular target stimulus (e.g., performance B), the features of that stimulus are weighted more heavily than the features of an alternative comparison stimulus (e.g., an earlier performance A). When options have unique positive features, the comparison process described above gives an advantage to option appearing in the second position. Increase of scores with serial position can be explained by the fact that participants of prestigious contests are usually high-level performers and thus have more unique positive than unique negative features.

Another explanation of ordering effects was suggested by Unkelbach, Ostheimer, Fasold and Memmert (2012), who used the idea of calibration (development of an internal scale during a judgment series) and empirically showed that judgments become more extreme (not necessarily more positive) later in a series of judgments. In their experiments judges evaluated the same good (poor) performances as more positive (negative) at the end of a sequence compared to the beginning. Therefore, the calibration explanation of serial position effects in evaluative judgments predicts a positive relationship between the running number of a participant and her result only for good performances. However, neither of the theoretical explanations rules out the possibility of a negative or a non-monotonous relationship between the running number and the evaluation, which makes empirical evidence from different settings (music, sport, job interviews, etc.) especially valuable.

Even though generally the research evidence indicates that later serial positions benefit from more positive evaluations, the issue was addressed by few naturalistic studies (Page & Page, 2010). The purpose of our study, which mostly replicates the studies of Glejser and Heyndels (2001) and Haan et al. (2005) using newer observational data from other music competitions, is to test the influence of a contestant's run-

*National Research University Higher School of Economics Kantemirovskaya St. 3, Saint-Petersburg, Russia, 194100. Email: eantipov@hse.ru.

†National Research University Higher School of Economics, Saint-Petersburg, Russia, 194100.

https://doi.org/10.1017/S1930297500006288 Published online by Cambridge University Press

ning number's rankings, using empirical data from several song contests. We are the first to compare the influence of an exogenous factor (the order in which singers perform) on the judgments of both professional experts and TV audience in Eurovision finals. Besides that, we have collected a very rich dataset of New Wave Song Contest results from 2005 to 2016, which is another relatively rare example of a contest in which a participant's order of appearance is determined by lot. The fact that New Wave consists of several rounds has allowed us to additionally test whether the order effect is weakened on the second and the third days compared to the first day of the contest.

It is worth mentioning that the Eurovision Song Contest has long attracted attention of academic researchers who looked at political and cultural determinants of the contest's results. Ginsburgh and Noury (2008) showed that the votes are driven by linguistic and cultural proximities between singers and voting countries. Spierdijk and Vellekoop (2009) established strong evidence for voting bias in the song contest on the basis of geography, even after controlling for culture, language, religion and ethnicity. Blangiardo and Baio (2014) used Bayesian hierarchical models and found moderate to substantial positive bias, which they explained by strong "cultural" similarities in language and history, and to a lesser extent to geographical proximity and migrations. Researchers also studied some voting biases which are not related to political and cultural voting. For example, Verrier (2012) found evidence for the influence of the mere-exposure effect on Eurovision voting by showing that contestants did better if they previously appeared in a semifinal that was seen by voters.

The Eurovision Song Contest provides a wealth of data for analyzing possible distortions arising from the cognitive biases in the evaluation of performance. However, the authors of all previous studies that made use of Eurovision data encountered serious data problems. Unavailability of data on the full ranking of countries outside the top-10 list created the problem of left-censoring (contestants that are given a rank greater than 10 all receive zero points, which does not mean they were all equally preferred by the jury or the audience). In addition, professional judges can be expected to be more impartial than televoters, which is why it would be useful to test the impact of various factors on professional jury's and TV viewers' judgments separately. However, the problem of the data used in earlier studies is the inability to isolate televotes from the votes of professionals in years when both televoting and jury voting contributed to the outcome of the competition. Due to the unavailability of such detailed data from the final rounds, Haan and his coauthors (2005) had to test for the difference between experts judgments and public opinion by taking advantage of the fact that some national finals of the Eurovision Song Contest are judged by a jury of experts, while others are decided by televoting. This may have introduced some sample selection bias, since the method of choosing a candidate for the European Song Contest is likely to be endogenous (if a country relies on televoting, this may indicate that the audience is considered more musically educated than the audience in countries where producers rely on professional opinion). In order to make the comparison perfect from the statistical point of view it would be desirable if in each country both the jury and the audience voted for the contestants, which was the case in 2009–2015 festivals.

2 Materials and Methods

We present the results based on the data from two song contests: the world-famous Eurovision Song Contest founded in 1956 and the New Wave, an influential international contest for young performers of popular music held annually since 2002. Despite a large number of different contests, these two have a sufficiently long history of publicly available results together with the sequence of participants, which is known to have been determined exogenously by lot in both contests at least in those years that were included in our dataset.

To increase openness, the organizers of the Eurovision Song Contest decided that from 2009 onwards, the detailed split jury and televoting results would be revealed, making it possible to disentangle the rankings by professional jury and televoters. Starting from 2014 they started to publish even more detailed data, containing information on how each jury member ranked each of the contestants. By using new detailed data from the final stage of the European Song Contest we not only provide additional evidence of ordering effects when judging music contests, but also avoid any statistical problems, because of the availability of complete professional jury and televoting results for each of the countries. Despite the presence of every single juror's ranking of contestants, the problem with the most recent Eurovision data is that starting from 2013 the running order of participant was determined by producers to make the show more spectacular.

Even though the opportunity to learn how each of the jurors voted appeared only in 2014, the practice of publishing separate rankings based on jury voting and televoting started in 2009, which makes the data from 2009–2012 when the running order was random and thus appropriate for the analysis of order effect. Therefore, we decided to use 2009–2012 data to analyze the influence of the running number on the resulting ranks according to professional jury and television votings. In 2009–2012 all countries used televoting and/or SMS-voting (50%) and five-member juries (50%), apart from San Marino which is 100% jury due to country size. 2009–2012 contests followed a standard format. First, the songs are performed in an order predetermined by lot. Second, there is a break of 15 minutes, in which viewers at home can decide on their vote. Television viewers can vote via the official application, telephone and/or SMS. These

https://doi.org/10.1017/S1930297500006288 Published online by Cambridge University Press

Table 1: Pearson correlation coefficients between the transformed running number and transformed ranks reflecting the voting results.

Correlate	Correlation	One-tailed p-value	Observations
Transformed jury rank	–0.140	0.081	101
Transformed televoting rank	–0.205	0.020	101

votes determine 50% of the outcome and are gathered by the European Broadcasting Union's (EBU's) voting partner. Another 50% of the outcome is determined by the professional jury. Each national jury consists of 5 music industry professionals. The jury members shall rank first their favorite song, second, their second favourite song, third, their third favourite song, and so on until their least favorite song, which shall be ranked last. Third, each country's representative announces the scores (1–8, 10 and 12) for the country's top 10 favorite songs (based on the combined professional jury and televoting ranking). Since only 12, 10 and 1–8 points are being given countries ranked outside of the top-10 do not receive points. The song which has received the highest number of votes shall be ranked first, the song which has received the second highest number of votes shall be ranked second and so on until the last song. The contestant who scores the most becomes the winner and performs his/her song once more.

The second dataset contains the results of New Wave — an international contest for young performers of popular music[1]. It has 3 contest days. The order of performance has always been determined by lot, which gave us 12 well-documented years of data, where the order of contestants was an exogenous determinant of their results. The key difference between New Wave and Eurovision rules is that at New Wave each of the 10–15 judges raises a card with his or her score immediately after each performance, which allows testing for ordering effects when such an approach, that mimics sports judging, is used instead of a ranking procedure.

3 Results

3.1 Evidence from the Eurovision Song Contest

In order to correctly identify the influence of the running number on the opinion of professional judges and amateur audience we need the running order of contestants to be exogenous. We utilized the fact that in 2009–2012 the running order of the Eurovision contestants was determined by lot, i.e. randomly, and, at the same time, split jury and televoting results were made available by the EBU. Since the number of competing countries was 25 in 2009–2011 and 26 in 2012 for the regression analysis we standardized ranks to the interval [0,100] using a slightly modified formula by Haan et al. (2005):

$$R_{it}^{trans} = \left(1 - \frac{R_{it} - 1}{n_t - 1}\right) \cdot 100 \quad (1)$$

where R_{it}^{trans} is the transformed rank of the i_{th} country in year t, R_{it} is the untransformed rank (rank 1 corresponds to earlier performance in the case of the running number and the best performance in the case of jury/televoting ranks) and n_t is the number of contest participants in year t. To make the data comparable across years the transformation was applied to all the ranks involved in the analysis: running order, jury ranks and televoting ranks. Note that *higher* values of the transformed jury and televoting ranks imply *better* performance, while a *higher* value of the transformed running number implies a *smaller* running number (earlier performance). The transformation leads to values that are very close in meaning to percentiles, but ensures that 100 always corresponds to the lowest rank (1) in year t, while 0 corresponds to the highest rank (n_t).

The Pearson correlation coefficients between the transformed running number and transformed ranks reflecting the results based on jury voting and televoting with one-tailed p-values are given in Table 1.

For televoting the order effect is significant (p=0.020): the later a contestant performs, the higher he/she is ranked by televoters. For professional jury the effect is significant at 10% level (p=0.081). These are weak signs of inefficiency in the decision making process of Eurovision voters.

Given that the correlation between transformed televoting and transformed jury ranks is 0.4005, a test of significance for the difference between the two dependent correlation coefficients using the Steiger's Z test results in z=0.6 (one-tailed p=0.275), which indicates that the order effect is insignificantly weaker for professionals than for amateurs.

3.2 Evidence from the New Wave Song Contest

The correlation between transformed rank and transformed running number was –0.188 (p<0.001). The correlation remained highly significant (r=–0.149, p<0.001) even after filtering out the 2013 data, where the correlation was suspiciously strong (r=–0.621, p<0.001). Parameter estimates of regression models with the transformed rank as the dependent variable are presented in Table 2.

In Table 2, model 1 shows the simple regression of the transformed rank on transformed running number. Model 2 checks whether the impact of running number on the result

[1]The contest's official website is http://newwavestars.eu/en/

https://doi.org/10.1017/S1930297500006288 Published online by Cambridge University Press

Table 2: Parameter estimates of linear regressions of transformed ranks on transformed running number (New Wave contest 2005–2016).

Model	(1) Transformed rank	(2) Transformed rank	(3) Transformed rank
Transformed running number	–0.186*** (0.0398)	–0.185*** (0.0478)	–0.161*** (0.0454)
Transformed running number*Round2		–0.0103 (0.0514)	
Transformed running number*Round3		0.00500 (0.0510)	
Constant	59.23*** (2.352)	59.23*** (2.356)	58.46*** (2.446)
N	576	576	541
R^2	0.035	0.036	0.023
adj. R^2	0.034	0.031	0.021

Note: Heteroscedasticity-robust standard errors in parentheses.
One-tailed significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

is stronger on the first day of the contest than on days 2 and 3 due to the possible consequences of judges being familiar with the contestants on days 2 and 3 as opposed to day 1. The coefficients in the second and third rows represent the interactions between running number and round, which appear to be quite small. In model 3 we have left participants who were first performers (running number=1) out of the sample to check whether the coefficient at the running number is a mere reflection of the difference between the first participant, who is compared to nobody, and all others. The effect of transformed running number is still quite strong.

4 Conclusion

Our study based on the New Wave contest data has shown that expert judgments are influenced by the contestant's running number – an exogeneous factor that probably does not influence the quality of output. Other things equal, the later a contestant performs, the higher he/she is expected to be ranked. This agrees with some previous studies. However, we have found ambiguous statistical evidence of such an effect in the Eurovision finals of 2009–2012, as well as no evidence of expert rankings being less influenced by order effects than televoting-based rankings, meaning that the result of one of a previous study, where experts were found to be unambiguously better judges of quality than televoters (Haan et al., 2005), does not generalize at least to some of the major song contests.

Despite ambiguous significance of order effects in the Eurovision song contest, some people may still consider them to be of practical importance. Taking into account that professional juries rank all songs based on the second Dress Rehearsal (the so-called Jury Final), it may be advisable to reverse the running numbers of participants during the Jury Final to weaken the ordering effect with the help of televoting. A similar technique is often used in survey research when options in a question are presented in a different order to each participant.

A limitation of our study is that the quality of songs may not be fully randomized despite the fact that a draw was used to determine the order of contestants. As Bruine de Bruin (2003) discusses, this may arise as later contestants may view earlier contestants' performances, which may increase their motivation. Thus, it is possible for song quality to increase with increasing running number. However, because of the subjectivity of quality in performing arts it is hard to isolate order effect from such a motivation effect.

Our study uses new data that became available recently, when the EBU decided to make the contest more transparent. We believe that the appearance of detailed Eurovision Song Contest results will not only increase openness and credibility of the contest, but will also stimulate research studies on Eurovision including modified replications of previous studies that were not able to account for the differences between professional and amateur voters because of the lack of such data at the time those studies were conducted. One of the directions for future research is studying the impact of personal characteristics of experts on their judgments (e.g., whether there are gender differences in the extent of judgmental bias). In addition, new data will serve as good empirical material for those studying optimal ways to aggregate preferences (Besson & Robardet, 2007). It would also be useful to study the significance of the ordering effects in other performing arts contests, where the running order is determined by a random draw. In addition, the inter-rater reliability has never been studied for song contests before. While in the case of sport competitions like gymnastics or figure skating a high concordance of scores given by different judges is ensured by the codes of points (Bučar, Čuk, Pajek, Karacsony & Leskošek, 2012; Leskošek, Čuk, Karácsony, Pajek & Bučar, 2010), the inter-rater agreement on the quality of art remains an open question.

https://doi.org/10.1017/S1930297500006288 Published online by Cambridge University Press

References

Besson, J., & Robardet, C. (2007). A new way to aggregate preferences: application to Eurovision song contests. In *Advances in Intelligent Data Analysis VII* (pp. 152–162). Springer.

Blangiardo, M., & Baio, G. (2014). Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models. *Journal of Applied Statistics*, *41*(10), 2312–2322.

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245–260.

Bruine de Bruin, W., & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes*, *92*(1), 91–101.

Bučar, M., Čuk, I., Pajek, J., Karacsony, I., & Leskošek, B. (2012). Reliability and validity of judging in women's artistic gymnastics at University Games 2009. *European Journal of Sport Science*, *12*(3), 207–215.

Glejser, H., & Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth Music Contest. *Journal of Cultural Economics*, *25*(2), 109–129.

Haan, M., Dijkstra, S., & Dijkstra, P. (2005). Expert Judgment Versus Public Opinion–Evidence from the Eurovision Song Contest. *Journal of Cultural Economics*, *29*(2), 59–78. http://doi.org/10.1007/s10824-005-6830-0.

Leskošek, B., Čuk, I., Karácsony, I., Pajek, J., & Bučar, M. (2010). Reliability and validity of judging in men's artistic gymnastics at the 2009 university games. *Science of Gymnastics Journal*, *2*(1), 25–34.

Page, L., & Page, K. (2010). Last shall be first: A field study of biases in sequential performance evaluation on the Idol series. *Journal of Economic Behavior & Organization*, *73*(2), 186–198.

Stewart, J. M., O'Shea, E., Donaldson, C., & Shackley, P. (2002). Do ordering effects matter in willingness-to-pay studies of health care? *Journal of Health Economics*, *21*(4), 585–599.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. http://doi.org/10.1037/0033-295X.84.4.327.

Unkelbach, C., Ostheimer, V., Fasold, F., & Memmert, D. (2012). A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes*, *119*(1), 103–113.

Verrier, D. (2012). Evidence for the influence of the mere-exposure effect on voting in the Eurovision Song Contest. *Judgement and Decision Making*, *7*(5), 639–643.