Available online at www.sciencedirect.com
SCIENCE DIRECT®
Acta Psychologica 118 (2005) 245–260
acta psychologica
www.elsevier.com/locate/actpsy

Save the last dance for me: unwanted serial position effects in jury evaluations

Wändi Bruine de Bruin *

Department of Technology Management, University of Technology, Eindhoven, The Netherlands and Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, USA

Received 8 June 2004; received in revised form 31 August 2004; accepted 31 August 2004

Abstract

Whenever competing options are considered in sequence, their evaluations may be affected by order of appearance. Such serial position effects would threaten the fairness of competitions using jury evaluations. Randomization cannot reduce potential order effects, but it does give candidates an equal chance of being assigned to preferred serial positions. Whether, or what, serial position effects emerge may depend on the cognitive demands of the judgment task. In *end-of-sequence* procedures, final scores are not given until all candidates have performed, possibly burdening judges' memory. If judges' evaluations are based on how well they remember performances, serial position effects may resemble those found with free recall. Candidates may also be evaluated *step-by-step*, immediately after each performance. This procedure should not burden memory, though it may produce different serial position effects. Yet, this paper reports similar serial position effects with end-of-sequence and step-by-step procedures used for the Eurovision Song Contest: Ratings increased with serial position. The linear order effect was replicated in the step-by-step judgments of World and European Figure Skating Contests. It is proposed that, independent of the evaluation procedure, judges' initial impressions of sequentially appearing candidates may be formed step-by-step, yielding serial position effects.

* Postal address: Department of Social and Decision Sciences, Carnegie Mellon University, Porter Hall 208, Pittsburgh, PA 15213, USA.
E-mail address: wandi@cmu.edu

0001-6918/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2004.08.005

PsycINFO classification: 2340
Keywords: Human judgment; Judgment and decision making; Sequential effects

1. Introduction

In many judgment tasks, options are presented in sequence. Consider, for example, the evaluation of job applicants, students' exams, apartments, and candidates of formal competitions such as the World Figure Skating Contest. In each of these contexts, judgments may be affected by the order of presentation, such that contenders of the same quality may receive a better rating in one serial position than in another. Such serial position effects could threaten the fairness of competitions that use jury evaluations, as well as the subsequent careers of the contestants (Ginsburgh & van Ours, 2003).

Researchers of judgment and decision making have paid relatively little attention to serial position effects on evaluations. Most preference elicitation studies present items jointly, at the same time. When experimenters do opt for sequential presentation, counterbalancing is typically used to deal with potential serial position effects. This method presents participants with different presentation orders, calculating the average judgment of each option across participants as well as serial positions. Doing so, it treats order effects as noise, leaving them unexplored.

Outside of the psychological laboratory, counterbalancing is often not a feasible strategy to deal with serial position effects. In many contests, for example, all jury members watch the sequentially appearing candidates in the same order. If all judges are vulnerable to similar serial position effects, then these may be amplified in their combined evaluations.

Presumably suspecting serial position effects, many formal contests have performers draw lots to determine their serial position. While randomization cannot reduce potential order effects, it does give candidates an equal chance at appearing in preferred serial positions. In that sense, randomization may be seen as improving the fairness of a competition that uses jury evaluations.

Fairness may be further increased by choosing the judgment procedure that is least likely to create order effects. Formal competitions often use one of two judgment procedures. Some contests require *end-of-sequence* judgments, made after all candidates have performed. With *step-by-step* procedures, each candidate has to be evaluated immediately after performing, before the next one takes the stage. Seemingly irrelevant variations across evaluation procedures may pose different cognitive demands, and possibly, affect the size and the direction of serial position effects.

When asked to make end-of-sequence judgments, judges may find it difficult to remember all performances. As the number of sequentially presented options increases, it becomes less likely that each of them will be recalled (Anderson, Bothell, Lebiere, & Matessa, 1998; Glenberg et al., 1980). Independent of the number of options, the probability of recall is typically higher for the very first and the very last

presentation, decreases for neighboring items that are further removed from the beginning and the end, and is "somewhat flat in intermediate positions" (Anderson et al., 1998, p. 366). In competitions, attempts to remember performances may be hindered by breaks and announcements. Prolonged distractions between items, and during the retention interval, often reduce serial position effects to the very first and the very last item (Glenberg et al., 1980).

Recall may affect evaluations if judges use a form of the availability heuristic (Tversky & Kahneman, 1973). Taking the degree to which an option is remembered as an indication of its quality, judges may give higher scores to performances they remember better. Presumably, this would benefit the very first and the very last performances in competitions using end-of-sequence evaluation procedures.

A survey conducted among members of the Society for Judgment and Decision Making (JDM) suggested that these experts also expected memory limitations to produce serial position effects in end-of-sequence evaluations (Bruine de Bruin & Keren, 2003a). When asked to predict which serial position would give a candidate a better chance of winning a hypothetical competition using an end-of-sequence procedure, their collective responses formed the serial position curve known from free recall experiments. Those who volunteered an explanation referred to the serial position effect in free recall. JDM members asked about the step-by-step procedure were less likely to expect candidates to benefit from performing in the first and the last few serial positions. Two-thirds of another group of JDMers answered the question "which procedure would, in your opinion, be least likely to produce order effects in the jury's evaluation of the candidates?" by selecting the step-by-step procedure.

Although using step-by-step judgments may reduce the burden on a judge's memory, it may pose other cognitive challenges, producing different serial position effects. For example, step-by-step processing forces judges to evaluate performances in one order, comparing each performance to earlier, but not to later ones. In such unidirectional comparisons, jury members may overweigh the unique features of each new, focal, performance (Tversky, 1977). If each sequentially presented option has positive unique features, it may seem better than previous ones, leading to higher scores with increasing serial position (Bruine de Bruin & Keren, 2003b; Houston, Sherman, & Baker, 1989). This direction-of-comparison effect is less likely to produce decreasing ratings in sequentially presented options with unique negative features (Bruine de Bruin & Keren, 2003b). This pattern may be explained by the finding that judges give more attention to and have better memory for negative than positive features (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Willemsen & Keren, 2002). As a result, unique negative features of previous items may be less likely to be forgotten, or ignored, compared to positive unique ones, when a new item appears.

Despite procedural differences, serial position effects due to direction of comparison are actually similar in step-by-step and end-of-sequence judgments of sequentially presented options (Bruine de Bruin & Keren, 2003b). Studies asking judges to revise a verdict step-by-step, after each piece of sequentially presented evidence, or end-of-sequence, after everything has been considered, also report similar order effects in both procedures (Hastie & Park, 1986; Hogarth & Einhorn, 1992). These

results suggest that end-of-sequence and step-by-step procedures may yield similar processing. That is, end-of-sequence judgments may be based on initial impressions that were formed step-by-step.

Step-by-step judgments may also affect the extremeness of judges' scores. The relative quality of the first option may not be evident until a second one has appeared (Moore, 1999). Judges who experience uncertainty when evaluating performances in low serial positions, may strategically use values near the middle of the scale. Doing so, they leave room to move upward or downward when evaluating later candidates. Judges using an end-of-sequence procedure should not face such uncertainty. The effect may not be completely eliminated, however, if judges make insufficient adjustments from the ratings they initially made step-by-step, as candidates performed (Tversky & Kahneman, 1973).

To date, serial position effects on jury evaluations have been examined in only a few formal international competitions. Across different finals of the Queen Elisabeth Competition for classical violin and piano, musicians performing on a later day received better end-of-sequence judgments (Flôres & Ginsburgh, 1996). Twelve finalists performed at a rate of two a day, with better scores being obtained by performances that were scheduled later in the week as well as later in the evening (Glejser & Heyndels, 2001). These serial position effects occurred in the evaluations made by a jury of 15 highly qualified experts.

Negative correlations between serial positions and final ranks were also reported for the 1973 World Championship in synchronized swimming and an amateur meet held in the same year (Wilson, 1977). In both competitions, final ranks were based on two rounds of performances, each judged by a different experienced jury. Because only two editions of these competitions were analyzed, it is unclear whether the results hold across synchronized swimming contests.

A more recent manuscript examined jury evaluations made for the Eurovision Song Contest, a popular music competition among artists representing different European countries. Scores increased with serial position, and more so when lay judges used televoting than when official juries used more formalized procedures (Haan, Dijkstra, & Dijksta, 2003). Over the years, the organizers of the competition have asked official juries to switch from end-of-sequence to step-by-step judgments.

This paper examines serial position effects on jury evaluations of the Eurovision Song Contest, comparing its end-of-sequence and step-by-step procedures, and of World and European Figure Skating Contests, which has consistently enforced step-by-step judgments. Specifically, this paper examines the following three hypotheses about serial position effects, which are not necessarily mutually exclusive:

Hypothesis 1: (a) End-of-sequence procedures result in relatively high scores for the very first and the very last items, reflecting serial position effects on free recall. (b) Such serial position effects should not occur in competitions using the step-by-step procedure, where performances are judged immediately.

Hypothesis 2: Competitions using either of the two procedures show increasing scores with serial position, due to direction-of-comparison effects.

Hypothesis 3: Competitions using either of the two procedures show the use of more extreme scale values with serial position, reflecting judges' uncertainty about their initial evaluations of earlier performances.

2. Study 1: Eurovision song contest

2.1. Contest procedure

The information described in this section summarizes information about the Eurovision Song Contest, collected by Walraven and Willems (2000). The European Broadcasting Union organized the first Eurovision Song Contest in 1956, inviting each of the associated countries to enter two original pop songs. Later competitions allowed only one entry per participant. Each country also contributed its own national jury, consisting mainly of lay people. To prevent nationalistic bias, judges have been prohibited from evaluating the performance representing their own country. Judges have been exposed to the dress rehearsal, allowing them to familiarize themselves with the entries.

Scores have been made public since 1957. At the end of most competitions since then, a spokesperson for each national jury called the presenters to announce the total number of points given to each performer by the national jury. Over the years, different scoring systems were used to arrive at these total scores. Through 1974, members of national juries followed various end-of-sequence procedures in making their final judgments. In 1957–1961, 1967–1970, and 1974, 10 judges in each national jury each gave one point to their favorite song. In 1962, individual judges awarded three, two, and one point to their top three songs. After calculating combined scores, each national jury then gave three, two, and one point to its members' three favorite songs. Judges in 1963 also used a form of rank-ordering, awarding 5, 4, 3, 2, and 1 point(s) to their top five. In 1964–1966, each member's three favorite songs received five points, the second four, and so on until the fifth, which received one point. In 1971–1973, two judges from each country judged each performance on a scale from one to five.

Starting in 1975, the organization switched to a step-by-step procedure, with ratings being collected immediately after each performance. Each jury member judged each performance on a scale from one to ten, with the top 10 songs across each national jury eventually receiving 12, 10, 8, and 7–1 points. Since 1998, viewers in most participating countries have been invited to cast a vote by calling a phone number corresponding to their favorite performance, at the end of the sequence. Televotes, or, if necessary, back-up jury evaluations, from each country were tallied and awarded 12, 10, 8, and 7–1 points to the 10 songs.

The analyses reported here use the data of the 47 editions of the Eurovision Song Contests held over the years 1957–2003, publicly available from fan sites

(http://www.kolumbus.fi/jarpen; http://www.songcontest.nl/Years). Fans and professionals have suggested a higher probability of winning for songs performed near the beginning or the end, countries that share culture with more national juries, the host, and songs performed in English (see http://www.kolumbus.fi/jarpen; Haan et al., 2003; Walraven & Willems, 2000; Yair, 1995). However, none of these analyses specifically compared serial position effects in Eurovision editions using end-of-sequence procedures with those observed in Eurovision editions using step-by-step procedures.

2.2. Results

2.2.1. Serial position effects on standardized scores (Hypotheses 1 and 2)

Because the scoring system has been changed over the years, participants' final scores were standardized within each contest. Unlike official jury members, televoters were never enforced to watch the entire Eurovision program (a concern raised by Walraven & Willems, 2000). Because lay judges who tuned in late may have been hesitant to vote for the songs that they missed, songs that were performed later in the sequence may have received more votes. A meta-analysis (see Rosenthal & DiMatteo, 2000) across the 1998–2003 competitions that included televoting identified the overall correlation between standardized scores and serial position, the z-value corresponding to the correlation, and the 95% confidence interval of that z-value. These meta-analysis statistics showed that standardized scores increased with serial position ($r = .23$, $z = .23$, 95% c.i. for $z = .04, .42$). Subsequent analyses excluded contests that used televoting from the set that used end-of-sequence procedures. The remaining data included 19 end-of-sequence and 22 step-by-step competitions used by formal national juries. The first set had an average of 15.5 (s.d. = 2.9) participants, and the latter 21.0 (s.d. = 2.3), showing a significant difference, $t(39) = 6.86$, $p < .001$.

Fig. 1 shows mean standardized scores for the different serial positions, across Eurovision Song Contests using end-of-sequence and those using step-by-step proce-

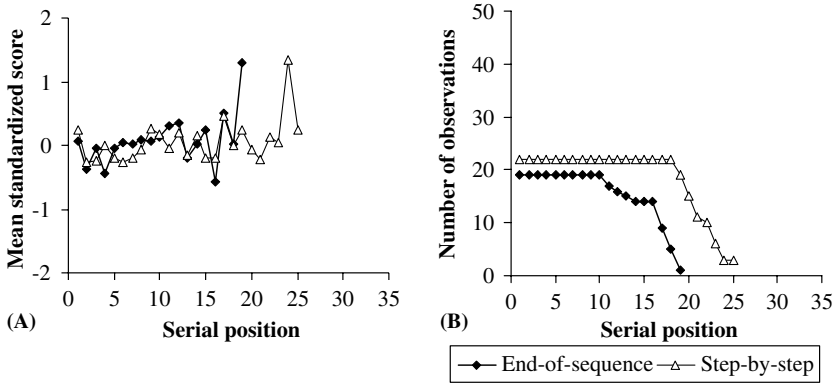

Fig. 1. Mean standardized scores (A) and number of observations (B) by randomized serial position, in Eurovision Song Contests using end-of-sequence and step-by-step procedures.

Table 1
Estimates of standardized scores in Eurovision Song Contest (R^2 = .25)

Predictor variables	*B*	*se*	*β*	*t*
Serial position	.02	.01	.11	2.26*
First	.36	.21	.08	1.72
Last	.21	.21	.05	.98
End-of-sequence (EOS)	−.05	.26	−.02	−.19
EOS* serial position	.00	.02	−.01	−.11
EOS* first	−.07	.31	−.01	−.23
EOS* last	.05	.31	.01	.17
Percent neighbors in jury	1.35	.89	.11	1.51
Home advantage	.39	.16	.08	2.38*
Finland	−.59	.23	−.13	−2.59*
France	.55	.20	.12	2.77**
Ireland	.92	.26	.19	3.59***
Portugal	−.51	.25	−.11	−2.03*
United Kingdom	1.17	.25	.27	4.62***

Note: Only dummy variables that are significant ($p < .05$) are presented in the table.
*$p < .05$, **$p < .001$, ***$p < .001$.

dures. Because the number of contestants varied across competitions, the later serial positions had fewer observations.

Table 1 shows the results of a linear regression examining serial position effects on standardized scores, with predictor variables including serial position, as well as performing as the very first and the very last in the sequence. The model also tested for a main effect of procedure (end-of-sequence vs. step-by-step), and an interaction of procedure with each of the order-effect variables (serial position, performing first, and performing last). It controlled for a potential home advantage benefiting the participant representing the host, as well as the proportion of national juries being from countries that shared land borders, and presumably more likely to share cultural tastes. Furthermore, the full model included dummy variables for the scoring system used, and for the country represented by the performer. Though conclusive data were missing for the language in which each song was performed, dummies for the country represented by the performer should control for most of this potential effect. Most of the contests used in the analyses enforced that performers sang in one of their native tongues, while the ones that did not had relatively few participants switch to English (Walraven & Willems, 2000).

Table 1 suggests no benefit of appearing first or last (Hypothesis 1), in addition to the increase of scores with serial position (Hypothesis 2). Dummy variables for scoring systems or end-of-sequence vs. step-by-step procedures were not significant. Scores were not significantly influenced by the proportion of national juries who were neighbors, but did increase if a performer's home country hosted that year's competition. Some countries proved more popular than others, with English-language countries UK and Ireland generally receiving better scores. France also systematically scored better than the rest, while Finland and Portugal did worse.

Dummy variables for the second, third, next to last and second to last serial positions showed no additional predictive ability, and neither did and dummy variables for each contest. There was no effect of adjusting serial position by dividing it by the number of candidates in each year's competition. A partial model, leaving out the predictor variables that were not significant, revealed similar estimates for the effect of serial position.

2.2.2. Serial position effects on extremeness of scores (Hypothesis 3)

The extremeness of the score given to each participant by each national jury was reflected by its absolute distance from the middle of that year's scale. Each participant's mean absolute distance value was calculated across national juries. For each of the individual 41 contests, a correlation was computed between participants' mean absolute distance values and their serial positions. As opposed to what was predicted by Hypothesis 3, a meta-analysis across these correlations showed significantly less extreme scores with serial position in end-of-sequence procedures ($r = -.12$, $z = -.12$, 95% c.i for $z = -.22, -.02$), and no significant effect in step-by-step procedures ($r = .00$, $z = .00$, 95% c.i. for $z = -.07, .06$).

The surprising result for end-of-sequence competitions was driven by the 16 that asked national juries to award points only to their one, three, or five favorites ($r = -.15$, $z = -.15$, 95% c.i. for $z = -.25, .-04$). In each of these cases, zero points were given to the rest. Uncertain judges may have reserved their few valuable points for later candidates. Doing so, they would give many extreme values of zero to earlier performances, producing a negative correlation between extremeness of scores and serial position. The three end-of-sequence competitions that, like all step-by-step competitions, allowed individual judges to evaluate each and every candidate on a rating scale showed no significant relationship between extremeness across scores and serial position ($r = .03$, $z = .03$, 95% c.i. for $z = -.20, .26$).

2.3. Discussion

These results provided no support for Hypothesis 1. Across end-of-sequence editions of the Eurovision Song Contests, there was no benefit to performing first or last. The serial position curve for standardized scores across end-of-sequence competitions did not appear like those known from memory research (Anderson et al., 1998; Glenberg et al., 1980). Scores for contestants in intermediate positions did not show a plateau, but, instead, there was an overall linear increase with serial position. Thus, memory limitations probably did not play a role in the reported results.

Rather, competitions using either of the two procedures showed increasing scores with serial position, suggesting support for Hypothesis 2. Changing procedures apparently did not prevent an unwanted serial position effect. Despite the relatively small effect size, such a serial position effect can have serious consequences for contestants' careers (Ginsburgh & van Ours, 2003), and pose a challenge to the fairness of the Eurovision song contest.

Previous experimental research (Bruine de Bruin & Keren, 2003b) provides a possible explanation for the reported results. Because performers appeared in sequence,

judges may have formed their initial impressions step-by-step. Each performance may have been judged in comparison to previous ones, with the unique features of that performance receiving more attention. Ratings may have increased with serial position because participants had mostly positive unique features, or because judges paid more attention to positive rather than negative unique features.

In addition to serial position effects, a threat to fairness was also suggested by the significant effects of control variables, indicating systematically better scores for the host, and songs performed in English by the UK and Ireland (see http://www.kolumbus.fi/jarpen; Haan et al., 2003; Walraven & Willems, 2000). Overall, France also received relatively high scores, while Finland and Portugal scored relatively poorly, suggesting systematic favoritism in national juries. Countries did not, however, systematically benefit from being judged by a jury consisting of a higher percentage of neighbors.

There was no indication that the extremeness of scores increased with serial position (Hypothesis 3), at least in end-of-sequence and step-by-step competitions that used a rating scale. End-of-sequence procedures that asked judges to give points to their favorite top one, three, or five produced less extreme scores over time, possibly because they gave the extreme value of zero points to many of the earlier performances. Overall, Eurovision judges may not have been more uncertain about judging earlier performances. Having witnessed the dress rehearsal may have helped judges to understand what quality to expect.

Because the end-of-sequence and step-by-step procedures used for the Eurovision Song Contest differ in more respects than just the timing of evaluations, caution is warranted when interpreting differences between these procedures. To make the procedures more comparable, end-of-sequence televoting procedures were not included in the analyses. Televoters who tuned in late may have voted only for performances they actually saw, favoring those in later serial positions (e.g., Walraven & Willems, 2000). However, formal national juries who viewed all performances produced similar serial position effects. Controlled experiments would be needed to determine whether similar cognitive demands may have produced the reported serial position effects in both procedures, and to identify ways to effectively reduce those demands. Study 2 examines whether the reported serial position effects hold in international figure skating competitions, using professional judges and step-by-step procedures.

3. Study 2: European and world figure skating championships

3.1. Contest procedure

Copies of the protocols from European and World Figure Skating Championships in 1994–2000 were obtained from the KNSB (the Dutch office of the International Skating Union). Unfortunately, protocols for the European Championships in 1995 and 1999 were missing, as were those for the second round in the 1999 World Championship for pairs and all 2000 World Championships. Thus,

the analyses reported here used results of the first round of 36 contests, and the second round of 32—held for men, women, and pairs, in European and World-level championships.

The two rounds in each international figure skating contest were (1) the Short Program (in which all participants performed the same figures) and (2) the Free Skating Program (in which figures were chosen by the participants). Serial position in the first round of each competition was randomized by drawing lots. Serial position in the second round was determined by the results of the first round. Based on the scores they obtained in the first round, figure skaters were divided into groups that determined the serial position in the second round. That is, the lowest placed group performed first, the next lowest second, and so on. The specific starting order within each group was determined by lots.

If the figure skating data show the linear order effects found with the Eurovision data, then this procedure lets the random draw for serial position in the first round affect scores in the first as well as the second round. To test this idea, the reported analyses examine effects of—serial position in the first round on scores obtained in that round, as well as scores obtained in the second round.

For each contest, nine professional judges from nine different countries were randomly selected from an international pool. Eligible judges received extensive training to achieve high inter-rater reliability, had years of world-level jury experience, and were continuously checked for nationalistic bias (Weekley & Gier, 1989). In each figure skating contest, evaluations made by each individual judge were publicly posted after each performance—thus followed a step-by-step procedure. In both rounds of each competition, judges awarded up to 12.0 points, using a maximum of 6.0 points for each of two dimensions. In the first round, points were given for the quality of the required elements, and for presentation, which includes the beauty of the routine. In the second round, points reflected technical merit, such as style and choreography, as well as artistic impression. Scores for the two dimensions judged in each round were typically highly correlated (Weekley & Gier, 1989).

3.2. Results

3.2.1. Serial position effects on standardized scores (Hypotheses 1 and 2)

To make analyses comparable to those of Study 1, participants' total scores were standardized within each competition. Fig. 2 shows the standardized scores obtained in the first and the second round, plotted by skaters' randomized serial position in the first round. Because the number of observations decreased with serial position, mean standardized scores for later serial positions may be more difficult to interpret. The number of participants averaged 25.8 (s.d. = 5.0) and 22.7 (s.d. = 3.3) in the first and second rounds, respectively.

Table 2 shows the results of a linear regression predicting standardized scores in the first and the second round. Predictor variables were performing first, performing last, and linear order of appearance, while controlling for a potential home advantage and being judged by a jury that includes a member of one's own nationality. Dummy variables were included for each performer's nationality. Predictions

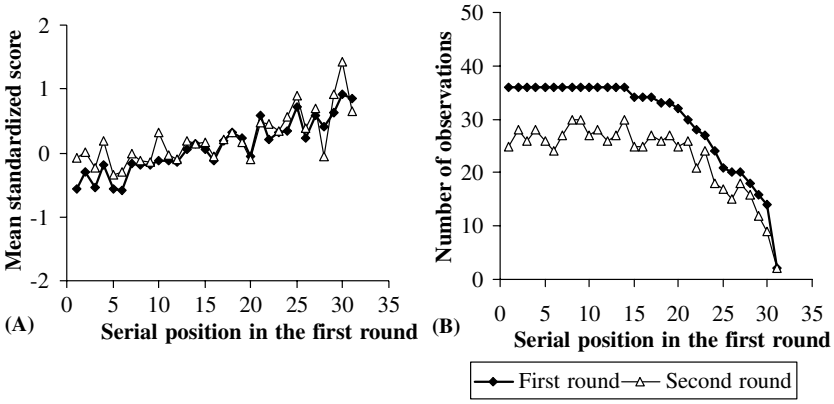

Fig. 2. Mean standardized scores (A) and number of observations (B) by randomized serial position in the first round, for figure skating contests.

Table 2
Estimates of standardized scores in the first and second round of World and European Figure Skating Contests

Predictor variables	First round (R^2 = .56)				Second round (R^2 = .47)			
	B	*se*	*β*	*t*	*B*	*se*	*β*	*t*
Serial position	.02	.00	.20	7.61***	.02	.00	.12	3.58***
First	−.05	.12	−.01	−.40	.18	.18	.03	1.02
Last	.00	.12	.00	.03	.24	.17	.04	1.44
Juror from own country	−.03	.05	−.02	−.59	.05	.07	.02	.68
Home advantage	.07	.12	.01	.57	.13	.16	.03	.87

Note: All dummy variables for the country represented by the performer are significant ($p < .05$).
*$p < .05$, **$p < .001$, ***$p < .001$.

of the standardized scores in the first round, presented in the left-hand column of Table 2, should be considered to examine hypotheses 1 and 2. Overall, appearing first or last did not provide explanatory power in addition to a linear order effect. As with the end-of-sequence and step-by-step judgments of Eurovision artists, figure skaters who appeared later were judged as better. Performers' scores did not show a home advantage, or benefits from being judged by a jury member with the same nationality. Dummy variables were significant for all participating countries ($p < .05$).

The right-hand side of Table 2 shows the results of a linear regression predicting standardized scores in the second round, from variables specific to the first round. They confirmed that the random draw for serial position in the first round predicted who received better scores in the second round.

In both cases, adding dummy variables for the second, third, next to last and second to last serial positions, showed no additional significant effects, and did not affect the reported pattern of results. Neither did adding dummy variables for each contest, or adjusting serial position by dividing it by the number of candidates in each year's competition. Separate partial models predicted scores obtained in the first

round, and scores obtained in the second round, using only the predictor variables that proved significant in the corresponding full model. They predicted similar serial position effects as the full models reported in Table 2.

3.2.2. Serial position effects on the extremeness of scores (Hypothesis 3)

The extremeness of scores was reflected in the absolute distance between the raw score given to each performer by each juror, and the middle of the scale. Each performer received a mean absolute distance value, averaged across jury members. For each of the 36 contests, a correlation was calculated between performers' mean absolute distance values and serial positions. Meta-analysis statistics (see Rosenthal & DiMatteo, 2000) showed that scores got more extreme with serial position ($r = .39$; $z = .41$, 95% c.i. for $z = .31$, .51), suggesting support for Hypothesis 3. Across the 36 competitions, correlations between the mean absolute difference from the middle of the scale and serial position were stronger in competitions with more contestants ($r = .34$, $p < .05$).

3.3. Discussion

As predicted by Hypothesis 2, linear order effects were detected in the step-by-step judgments of world-level figure skaters: Candidates performing later received better ratings. Possibly, judges focused on the positive unique features of each new appearing figure skater (Bruine de Bruin & Keren, 2003b). As predicted by Hypothesis 1, there were no additional benefits to performing first or last. These results were expected with the end-of-sequence procedure, where judges may base their final rating of a performance on how well they remember it.

Whatever the reason for the observed linear order effects, they threaten the fairness of jury evaluations. The tradition of letting skaters who received better scores in the first round perform later in the second round may, inadvertently, have aggravated serial position effects. The analyses showed that a random draw for serial position in the first round may affect a candidate's scores in the first *and* in the second round. Because final scores are a sum of both scores, a lucky draw may determine the winner of an international figure skating contest.

If the International Skating Union is concerned with the fairness of its procedures, it may be wise to randomly determine starting order in the first round, and let figure skaters perform in reverse order in the second round. An alternative solution would be to let order of appearance in the second round be determined by a new random draw.

Possibly due to world-level figure skating judges' extensive training (Weekley & Gier, 1989), their scores showed no systematic bias for figure skaters representing their home country or the host. However, judges' training for high reliability did not prevent serial position effects. Such training may actually have encouraged new judges to replicate the serial position effects produced by experienced judges. The tradition of letting skaters who received better scores in the first round perform later in the second round may further confirm the expectation that scores should increase with serial position.

Judges' extensive training also did not stop them from saving more extreme scale values for performances in higher serial positions, as predicted by Hypothesis 3. Possibly, judges were collectively uncertain about the quality of earlier performers, especially with larger number of competitors. Because scores increased with serial position, judges must have felt more hesitant about giving performers in earlier serial positions high extreme scores, compared to low extreme scores. Note that a linear order effect does not necessarily imply that more extreme scale values were reserved for later serial positions, as was observed in the Eurovision data. Starting with low extreme scale values, and ending with high extreme scale values, ratings could increase with serial position, without creating a relationship between serial position and the absolute distance of scores from the middle of the scale.

If the observed scale use effects were indeed due to uncertainty about judging earlier candidates, it may be reduced by watching practice rounds. Eurovision judges allowed to listen to dress rehearsals, did not show more extreme scores with serial position.

4. General discussion

End-of-sequence and step-by-step procedures used in the Eurovision Song Contest showed linear order effects of a similar pattern and a similar magnitude: Scores increased with serial position. A similar linear order effect was also found in the step-by-step judgments made for international figure skating competitions. These results replicate the linear pattern reported for synchronized swimming contests using step-by-step judgments (Wilson, 1977) and a classical music competition using end-of-sequence judgments (Flôres & Ginsburgh, 1996; Glejser & Heyndels, 2001). Thus, it appears that performers who appear later may receive more favorable evaluations, independent of whether the judgment procedure is end-of-sequence or step-by-step.

These results suggest that the serial position effects on end-of-sequence judgments were not due to memory effects. If that were the case, step-by-step judgments should not have shown the same pattern. Moreover, the linear serial position curve for standardized scores (Fig. 1) did not resemble the V-shaped serial position curves on serial recall (e.g., Anderson et al., 1998). Combined, these results seem to support Hypothesis 2, but not Hypothesis 1.

The similarity of the linear order effects in both procedures suggests that a similar evaluation process may have been applied to both. As with tasks involving sequential judgment of information about one option (Hastie & Park, 1986; Hogarth & Einhorn, 1992), judges in end-of-sequence procedures may have reduced the burden on their memory by forming initial impressions of candidates step-by-step, as they appeared. Though end-of-sequence procedures theoretically allowed judges to adapt their scores later, initial step-by-step impressions may have been resistant to change (Bruine de Bruin & Keren, 2003b). Judges who attempted to change initial impressions may have made adjustments that were insufficient (Tversky & Kahneman, 1973).

One possible explanation for the increasing linear order effects was tested in this paper. Hypothesis 3 predicted that judges would feel uncertain about how to evaluate earlier performances, for lack of comparison material. To be safe, they would safe more extreme scale values for later candidates, especially higher ones that could determine the winner. Well-trained figure skating judges showed this pattern, but Eurovision judges did not. The latter may not have experienced uncertainty when judging earlier candidates, because they heard the dress rehearsal. Eurovision scores nevertheless increased with serial position, suggesting that uncertainty about judging earlier candidates may not be the only explanation for the reported linear order effect.

An additional (not mutually exclusive) explanation is based on direction of comparison effects. Watching a sequence of performances, each new one may become the most salient. When making relative judgments, judges may have emphasized the unique features of the salient alternative (Tversky, 1977). Looking for the outstanding qualities of a winner, jury members may have noticed that the first figure skater made an impressive pirouette, the second an extraordinary double axle, and the third a breath taking choreography. Thus, positive unique features may have received more attention than shared ones, and made candidates seem better than earlier ones (Houston et al., 1989). Controlled experiments found increasing linear order effects due to direction of comparison in options with positive unique features, using step-by-step and end-of-sequence judgments (Bruine de Bruin & Keren, 2003b).

Finally, the reported linear order effects may also have reflected an actual increase in performance quality. Having seen others perform may have increased performers' goals or their achievement motivation, and, hence, their performance (for a review of the goal setting literature, see Locke & Latham, 1990). Effects of serial position on actual performance could not be tested in the present data sets, because there was no objective measure of performance quality.

Controlled experiments would be needed to tease out which of these explanations played a significant role in the reported order effects. Whatever the underlying mechanism, the present results suggest that linear order effects will occur whenever candidates appear in sequence—threatening the fairness of competitions. Randomization alone will let chance decide who performs later—and, consequently, gets a higher probability of winning. More research is also needed to examine whether serial position effects may be reduced by changing contest procedures, by training jury members to use the rating scale consistently over time, and teaching them to avoid unidirectional comparisons.

If possible, judgments of the same candidates should be made in different orders. In competitions that ask performers to make multiple appearances, the first could use a randomized order, the second its reverse, and so on. Similarly, teachers' grades may be fairer if they grade exams by question, using a different random order for each. In the absence of randomization, judgments may be affected by systematic factors that determine order of presentation. Candidates may then use their knowledge of the reported results to their advantage. When scheduling a job interview, for example, they may be better off booking the last slot in the sequence, keeping in mind The Drifters' 1961 hit song "*save the last dance for me*."

Acknowledgments

This research was made possible in part through support from the Department of Technology Management at the Eindhoven University of Technology, and the Center for Integrated Study of the Human Dimensions of Global Change, which has been created through a cooperative agreement between the National Science Foundation (SBR-9521914) and Carnegie Mellon University. I am indebted to Gideon Keren, as well as Jan van Bolhuis, Cobie Bruine de Bruin, Bruno Heyndels, Aïda Hordijk, Chris Snijders, and two anonymous reviewers for comments and advice in different stages of this project. Any remaining mistakes are my own.

References

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*, 341–480.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.

Bruine de Bruin, W., & Keren, G. (2003a). [Experts' predictions of serial position effects in competitions using end-of-sequence and step-by-step procedures]. Unpublished raw data.

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

Flôres, R. G., Jr. & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? The Statistician*, 45*, 97–104.

Ginsburgh, V. A., & van Ours, J. C. (2003). Expert opinion and compensation: Evidence from a musical competition. *American Economic Review, 93*, 289–296.

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*, 109–129.

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., Gretz, A. L., Fish, J. H., & Turpin, B. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 355–369.

Haan, M., Dijkstra, G., & Dijksta, P. (2003). Expert judgment versus public opinion—evidence from the Eurovision Song Contest. Unpublished Manuscript, University of Groningen.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*, 258–268.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief adjustment model. *Cognitive Psychology, 24*, 1–55.

Houston, D. A., Sherman, S. J., & Baker, S. M. (1989). The influence of unique features and direction of comparison on preferences. *Journal of Experimental Social Psychology, 25*, 121–141.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Moore, D. A. (1999). Order effects in preference judgments: Evidence for context dependence in the generation of preferences. *Organizational Behavior and Human Decision Processes, 78*, 146–165.

Rosenthal, R., & DiMatteo, M. R. (2000). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality & Social Psychology Review, 5*, 296–320.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.

Walraven, H., & Willems, G. (2000). *Dinge-dong. Het Eurovisie Songfestival in de twintigste eeuw*. Amsterdam, the Netherlands: Forum.

Weekley, J. A., & Gier, J. A. (1989). Ceilings in the reliability and validity of performance ratings: The case of expert raters. *Academy of Management Journal, 32*, 213–222.

Willemsen, M. C., & Keren, G. (2002). Negative-based prominence: The role of negative features in matching and choice. *Organizational Behavior and Human Decision Processes, 88*, 643–666.

Wilson, V. E. (1977). Objectivity and effect of order of appearance in judging of synchronized swimming meets. *Perceptual and Motor Skills, 44*, 295–298.

Yair, G. (1995). 'Unite unite Europe'. The political and cultural structures of Europe as reflected in the Eurovision Song Contest. *Social Networks, 17*, 147–161.