Luck of the Draw:
The Phenomenon of Serial Order Effects on Evaluation of Competitions©

Angela K. Antony, Harvard University

Introduction

Many formal and informal competitions hinge upon individuals or groups (judges) engaging in systems of evaluating candidates in successive order (competitions).

Does the order of presentation influence evaluations of these candidates?

If so, such "serial order effects" – defined as differences in evaluation resulting solely from the order of appearance—would compromise the integrity of competitions ranging from college admissions to Olympic sports.

Background Research

The appearance of stimuli in sequential order has shown demonstrated effects on memory (Murdock, 1962) perception (Asch, 1946) , attachment (Jaynes, 1957), persuasiveness (Knower, 1936), and preference (Carney & Banaji, 2008)—the vehicles through which people make competitive evaluations. However, almost no previous experimental research was conducted on the isolated effect of order on candidates' scores and evaluations in competitions.

Field Studies

A number of studies have investigated the impact of order on evaluation through observational studies of competition outcomes (Bruine de Bruin, 2005; Flôres & Ginsburgh, 1996; Page & Page, 2008; Wilson, 1977). Overall, each found a positive, linear trend between scores and serial position in the competition, suggesting that later serial positions earned higher scores.

Online vs. Retrospective Evaluation

Scores taken immediately after each performance in a competition ("on-line") were found to increase linearly with serial position, while those taken at the end of a competition ("retrospectively") mirrored the U-shaped free recall serial position curve (Redlawsk, 2001). These markedly different patterns suggest that the method of evaluation alone may impact the competition outcome.

Method

The Current Research

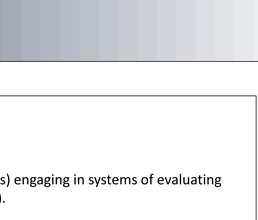
The role of order on evaluation was explored using a controlled experiment featuring 12 contestants from an actual singing competition.

The current research builds upon past work in the following ways:

- Studying the isolated effect of order using a controlled experiment (vs. observational)
- Studying a general audience evaluation rather than expert judges (vs. Plessner et al., 1999)
- Comparing on-line and retrospective evaluation of the same stimuli (vs. Arieli, 1998)
- Analyzing preference using a sequence greater than 2 (vs. Carney & Banaji, 2008)

Overall Vote Pages

Condition 1, Experiment 1 and 2



Experiment 1

[*n= 82, ages 18-52 (M = 20.29, SD = 4.75)]*

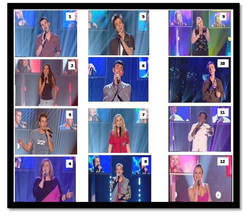
Materials

Twelve, 25-second video clips from the singing performances on the "Top 12" episode of the televised talent competition *Pop Idol* (Season 2) were the units of judging in this experiment.

Procedure

Each participant viewed one of four orders of the 12 video clips, and was asked to provide a rating on a scale from 1 (*extremely bad*) to 10 (*extremely good*) after each clip. After rating all performances, participants viewed a page with individual screenshots of all twelve performers—labeled 1-12 to match the sequence in which contestants were presented—and were asked to give one final, winner vote (see "Overall Vote Pages").

Condition 2, Experiment 1



Conditions

Participants were divided equally across four counterbalanced orders.

Condition 1:	1 → 12	**(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)**
Condition 2:	12 → 1	**(12, 11 ,10, 9 ,8 ,7, 6, 5, 4, 3, 2,1)**
Condition 3:	6 → 1, 12 → 7	**(6, 5, 4, 3, 2, 1, 12, 11 ,10, 9 ,8 ,7)**
Condition 4:	7 → 12, 1 → 6	**(7, 8, 9, 10, 11, 12, 1 ,2 ,3, 4, 5 ,6)**

Analyses conducted on the stimuli showed that no one person or performance stood out as exceptionally good or bad.

Experiment 2

[*n= 29, ages 18-22 (M =* 19.38, *SD =* 1.35*)]*

Experiment 2 utilized the same materials and procedure as Experiment 1, adding only one element: a passive viewing of all 12 performances at the start of the experiment. The goal of this was to give judges an overall sense of the quality of the talent pool before beginning the evaluation process. The stimuli were then scored on-line and awarded a retrospective winner vote in two orders: Condition 1: 1 → 12 and Condition 2: 6 → 1, 12 → 7.

Results

Experiment 1

On-line ratings. A repeated measures ANOVA revealed a highly significant serial position effect for order on on-line scoring. Scores had statistically significant linear ($F(1, 78) = 7.70$, $p = .01$), quadratic ($F(1, 78) = 45.46$, $p = .001$), and cubic ($F(1, 78) = 5.24$, $p = .03$) effects.

A visual inspection of the data indicated that ratings peaked in a fairly regular, concave-down quadratic pattern within each four positions (see Figure 1).

This pattern was consistent across four different arrangements of the 12 stimuli, suggesting the pattern is truly one of item position and not excellence of performer (see Figure 2).

Winner choice. The effect of serial position for retrospective voting roughly mirrored the U-shaped serial-position curve (see Figure 3)— approximately the reverse of the pattern observed for on-line scores. An exact binomial test confirmed this observation, revealing that the bottom-third of scored positions consistently received the highest number of votes ($p = .03$).

Table 1 (*top*) and Table 2 (*bottom*).

Scoring Outcomes by Condition (Exp 1)

	WINNER (1st place)			LOSER (12th place)		
ORDER	NAME	POSITION	SCORE	NAME	POSITION	SCORE
1	Michelle	9	6.08	Chris	12	3.5
2	Susanne	7	5.45	Michelle	4	3.35
3	Brian	2	5.30	Roxanne	8	3.65
4	Michelle	3	6.40	Kim	10	4.30

Voting Outcomes by Condition (Exp 1)

	WINNER (1st place)			LOSER (12th place)		
ORDER	NAME	POSITION	VOTES	NAME	POSITION	VOTES
1	Chris	12	6	Michelle*	9	0
2	Susanne	7	4	Brian*	8	0
3	Roxanne	8	6	Chris/Michelle/Brian	7, 10, 2	0
4	Chris	6	5	Kristy/Marc	7, 9	0

*More than 3 people tied for this position

Individual performer analysis. A Pearson correlation conducted on each clip's average score against that item's total number of votes revealed a significant negative correlation between scores and votes ($r(11) = -.69$, $p = .01$).

Overall, Michelle's performance wins by scores ($M = 5.23$, $SD = .44$), but ties for the fewest overall number of votes (1 out of 82 votes). Roxanne wins by overall votes (17 out of 82 votes) with Chris in second place for votes (14 out of 82 votes).

Experiment 1

Figure 1.

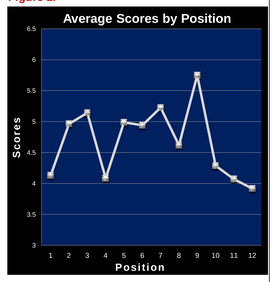


Figure 2.

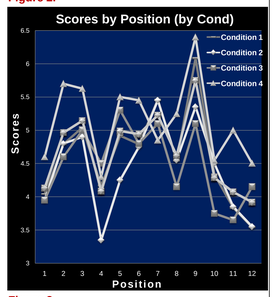
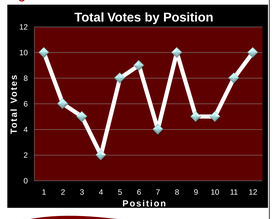


Figure 3.



Experiment 2

On-line ratings. A repeated measures ANOVA revealed a highly significant serial position effect for order on on-line scoring. Scores had statistically significant quadratic ($F(1, 27) = 11.64$, $p = .001$) and cubic ($F(1, 27) = 11.94$, $p = .001$) effects.

A visual inspection of the data indicated that scores across the 12 positions significantly peaked within each four positions, roughly mirroring the pattern observed in Experiment 1(see Figure 4 & Figure 5).

Figure 4.

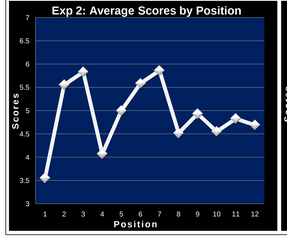
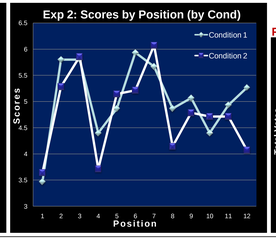


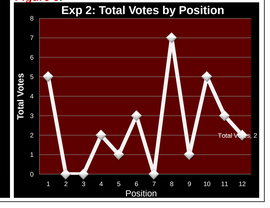
Figure 5.



Experiment 2

Winner choice. The effect of order on retrospective voting in Experiment 2 was again the reverse of the pattern observed for on-line scores (see Figure 6). An exact binomial test revealed that the bottom third of scored positions received a significantly higher number of votes (p =.001) and vice-versa (p = .01).

Figure 6.



Conclusion

Scores peaked at every 3rd/4th position, regardless of the specific content.
The position immediately after was consistently disadvantaged.

Why is three the magic number?

There is a substantial body of research suggesting that the working memory capacity for large, meaningful units of memory ("chunks") is three, and at a maximum four, items (Cowan, 2001; Golbet & Clarkson, 2004). The evaluation drop immediately after these positions is likely the result of reaching a memory capacity limit, leading the rater to revert to their anchor score. This attentional capacity limit is seen across the board: in developmental research, non-human animals, and cross-culturally.

Consider the world around you.

Cultural reflections of this psychological threshold are ubiquitous. Trichotomies occur in colloquial sayings ("three cheers," "third time's the charm," "one, two, three, go"), sports (baseball's three strikes and three outs; the definition of a "hat-trick"), and major religious symbols (the holy Trinity in Christianity) just to name a few (Dundes, 2008). Thus, it is no surprise that segmentation into threes also regularly occurs in competitive contexts—consider the traditions of bronze, silver, and gold medals in the Olympics, and *cum laude*, *magna cum laude*, and *summa cum laude* in the Latin honorary system.

~~~

Within this psychological context, it is reasonable that judges may subconsciously rank items in sequences of three—particularly when items do not deviate greatly in quality—thereby resulting in the specific, quadratic on-line scoring pattern observed.

This finding is aligned with previous work; the positive valence may confirm Bruine de Bruin & Keren's (2003) "direction-of-comparison" mechanism, and the linear scoring trend observed in numerous field studies may actually occur through incremental increases in sets of three positions, with an overall positive linear trend across the series.

So which is better? On-line scoring vs. retrospective voting. Neither scoring nor voting produced a consistent winner across orders, so it is possible that the true error is not contradictory methods of evaluation, but in the ambiguous title of "winner"—scoring and voting could be used intentionally to determine winners based on different criteria (such as voting to find "the whole package" in *Idol*, versus determining technical excellence with scores in the Olympics).

Suggested elimination techniques. Since on-line scores and retrospective votes demonstrated reversed effects due to serial order, a competition construction combining both methods of evaluation –such as a 50-50% weighted average of ranking by score and ranking by votes—could reduce or negate the effect of order bias.

Summary Discussion

Serial order effects are a bias that may challenge the fairness of both formal and informal competitions—from college applications to Olympic sports, Broadway auditions to beauty pageants. The current research explored the role of order on evaluation using a simulated, twelve-contestant singing competition.

On-line scores displayed a highly significant quadratic effect of order (see Figure 1), which suggested that raters awarded increasing scores in sequences of three positions, regardless of the content appearing in those positions. This specific scoring pattern was observed even when raters had prior knowledge of how each item compared to the entire talent pool (Experiment 2). It is proposed that this finding is linked to the upper limit of 3 to 4 items in "chunking" working memory capacity (Cowan, 2001).

The current research also investigated differences between on-line scoring and a retrospective "winner" vote for the same stimuli, finding that the outcome of the competition is decisively influenced by the method of evaluation. Results indicated that the positions (and individuals) that scored highest received the fewest votes, and vice-versa. In some instances the single lowest-evaluated performer of the competition was determined the winner by the other method of evaluation (Table 1 & 2).

Overall, these findings suggest that the outcomes of real world competitions may sometimes be decided by the competition construction—performance order and method of evaluation—rather than by the content of individual performances.

References

Ariely, D. (1998). Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations. *Journal of Behavioral Decision Making, 11*(1), 19-45.

Asch, S. E. (1997). Forming impressions of personality. In C. E. Hughes, D. Ariely & D. A. Eckerman (Eds.), *Background readings for the joy of experimental psychology (3rd ed.).* (pp. 48-72). Dubuque, IA, US: Kendall/Hunt Publishing Company.

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.

Carney, D., & Banaji, M. *First is best*. Unpublished manuscript.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-185.

Dundes, A. (2008). *The book of threes: The number three in American culture.* Retrieved 3/1, 2008, from http://www.threes.com/cms/index.php?option=com_content&task=view&id=15&Itemid=2

Flores, R., & Ginsburgh, V. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician*, *45*(1), 97-104.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four... or is it two? *Memory, 12*(6), 732-747.

Jaynes, J. (1957). Imprinting: The interaction of learned and innate behavior: II. the critical period. *Journal of Comparative and Physiological Psychology, 50*(1), 6-10.

Knower, F. H. (1936). Experimental studies of changes in attitude. II. A study of the effect of printed argument on changes in attitude. *The Journal of Abnormal and Social Psychology, 30*(4), 522-532.

Murdock, B. B. J. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482-488.

Page, L., & Page, K. (2008). Last shall be first: A field study of biases in sequential performance evaluation on the Idol series. Retrieved *March 1, 2006,* from *http://www.wbs-research.co.uk/pages/lionel/idol.pdf*

Redlawsk, D. P. (2001). You must remember this: A test of the on-line model of voting. *Journal of Politics, 63*(1), 29-58.

Wilson, V. E. (1977). Objectivity and effect of order of appearance in judging of synchronized swimming meets. *Perceptual and Motor Skills, 44*(1), 295-298.
~~~