When is a talent contest not a talent contest? Sequential performance bias in expert evaluation*

Alan Collins[†1], Jordi McKenzie[2], and Leighton Vaughan Williams[1]

[1]Nottingham Trent University, UK
[2]Macquarie University, Australia

January 28, 2019

Abstract

This study extends earlier work identifying sequence order biases in contest outcomes determined solely by popular voting. Results for different contest evaluation formats are empirically scrutinised, where both expert panel scoring and popular voting determine contest ranking. Forms of sequence order bias exist separately in the expert panel voting even though they are undertaken after each individual performance, as well as in the popular vote at the end of the contest. We suggest that the biases observed in the expert voting can be explained as a type of 'grade inflation'.

Key words: voting, performance evaluation, sequence order bias, expert judgement
JEL classifications: L82, Z10

[*] We thank the editor and an anonymous reviewer for comments that have helped to improve this paper. The authors are responsible for any errors.

[†] *Corrresponding author.* Email: alan.collins@ntu.ac.uk.

1 Introduction

Among other findings, Page and Page (2010) identify sequence order bias in the *Idol* televised singing contest where rankings are determined by popular voting at the end of each contest (episode). Later performing singers disproportionately gained higher popular vote shares, which they define as a 'recency' bias. Further, they find evidence of a small 'primacy' effect favouring the first performer.[1] Both types of biases implicitly relate to memory given the audience vote is after completion of each contest.

In principle, such sequence order effects should not feature in contests evaluated by expert panel judgement scoring on criterion-based terms. Such bias should also be minimised because panel scoring takes place after each performance. Nevertheless, using data derived from televised 'live' dance contest outcomes, we find evidence that there is sequential (J-curve) bias in both judges' evaluations and combined (television audience and judges) evaluations.

It is suggested that biases in judges' evaluations are 'non-designed' and could emanate, for example, from being influenced by close proximity to increasing studio audience acclaim and euphoria as each contest (episode) moves to conclusion.[2] In this sense, the sequential bias observed in the judges' scoring can be described as a type of 'grade inflation'. Even so, we do observe some evidence that first contestants within an episode are generally stronger performers, which could be the result of deliberate producer behaviour mindful of ratings and potential channel switching by the television audience.

2 Data

Individual expert panel score data and combined popular poll and judge rankings for the UK BBC television *Strictly Come Dancing* contest and the US spin-off on NBC called *Dancing with the Stars* were collected from television station (BBC and NBC) websites and Wikipedia pages. These contests pair a celebrity amateur dancer with a professional dancer. In the UK contest the judges' scores for each dancing couple are added together to give a total out of 40 (with four judges each able to give a maximum of 10 points for a given performance).

Once all couples have danced, they are ranked according to their total judges' scores with points awarded to each dance couple determined by their ranking. The number of points awarded is based on the number of dancing couples remaining in the competition. If eight

[1] Additional evidence of primacy and recency effects can be found in, for example, Mantonakis et al. (2009).

[2] A number of studies have considered crowd-based sources of bias entering sporting contests. Examples include Dowie (1982), Nevill et al. (2002), and Lenten et al. (2018).

couples remain, the couple at the top of the leader board receive eight points, the second seven points, etc.

Following this stage, it is open for the public to vote for their most preferred dancing couple. Since 2016 they could vote both by phone or online. The points from the judges' votes are then combined with points received via the public vote. Thus, if there were eight couples left in the competition and a couple came top in the judges' scoring and second in the public vote they would have accumulated eight points plus seven points, making a total of 15 points overall.

The public need to register a BBC account to vote online and may vote online (using that account) up to three times in each contest (excluding the Grand Final). People can also vote multiple times by phone. The lowest two scoring couples face a dance off where one is eliminated based on the expressed preferences of the judges. The head judge has the final deciding vote in the event of no unanimous decision by the other judges.

From Season 3 in the US spin-off, the voting system has been based on vote shares. The judges' vote share (the percentage of the total number of points awarded to all couples) and the public vote share for each couple are added together. The couple with the lowest combined total is eliminated from the show.

The BBC and NBC do not reveal the full popular vote data but only identify the lowest (two) scoring dancing couples. In order to focus more clearly on sequence order effects we examine only those episodes featuring one dance per couple. From the UK show, data was based on 15 seasons (over the period 2004 to 2017) each with typically 8 to 10 usable (single dance per couple) episodes. From the US show, data was extracted from 25 seasons over the period 2005 to 2017 (typically based on two seasons per year unlike the UK comprising one season per year) each with up to seven usable episodes.

3 Empirical analysis

Our first objective is to examine whether sequence order of contestants plays a role in elimination outcomes within a given episode. We follow Page and Page (2010) and construct a 'bias' variable based on the bottom two (or one for some of the US data) contestants eliminated. Although we do not observe audience votes, we do observe the judges' scores that provide additional insight about sequence order effects. For comparative purposes, we also construct a bias measure based only on the judges' bottom score(s).

To construct the sequence order bias variable, we compare the observed frequency of instances where a contestant featured in the bottom two (or one) positions against that which would have occurred randomly vis-à-vis sequence order position. More formally, for a given episode with N (remaining) contestants and $k \in \{1, 2\}$ achieving the lowest combined or judges' vote, we define the following measure of bias

$$E(Bias_{s,k,N}|s) = \frac{\sum \mathbb{1}(s = \text{safe})}{T_{k,N}} - \left(1 - \frac{k}{N}\right), \tag{1}$$

where s is the sequence order within an episode and $T_{k,N}$ is the total number of episodes of type $< k, N >$. In this equation, the first term refers to the actual frequencies that each sequence position is observed to be safe, while the second term reflects the theoretically random probabilities.

Figure 1 provides preliminary evidence that a J-curve effect exists in both the UK and US samples. Specifically, the first contestant in the sequence has either a positive (or small negative) bias, followed by a series of negative biases for low order sequences, and positive biases for contestants later in the sequence. Most notable, we observe a J-shape for *both* the combined *and* judges' scores, which suggests in addition to potential 'primacy' and 'recency' effects there is an additional sequence ordering bias arising from the judges' scores.

Tables 1 and 2 present econometric specifications of this relationship, where the dependent variable 'bias' is defined in Equation 1. Again, we follow Page and Page (2010) in construction of these models and consider both fixed effects (FE) and random effects (RE) specifications. As Page and Page (2010) describe in detail, the RE model is appropriate if sequence ordering is random, whereas the FE model is appropriate if ordering is non-random (e.g. producer imposed).

In all models there is supportive evidence of J-curve effects (i.e. positive 'first' and 'order' effects). Again, this is true for *both* the combined and judges' votes.[3] There is also evidence that 'order' effects are stronger in the combined data, which suggests the audience strengthen the J-curve effect. Although we do not observe this in the 'bottom one' US data, this does not imply that there are no such effects, merely that thy do not induce additional sequence biases. This would be the case if, for example, there was perfect correlation between the judges' scores and audience votes rankings.[4]

The parameter estimates across the FE and RE models are similar in both the UK and US specifications. The Hausman test suggests that the RE model is correct for both the UK and US data (with a single exception in one US specification). This suggests that ordering is random and the RE model is appropriate, which is consistent with Page and Page (2010).

[3] We investigated whether removing the first two seasons of the US series influenced results in any manner given the change in voting procedures but did not find differences in any result.

[4] Across all UK seasons, the correlation between final-place rank (based on combined scores) and average dance score rank (based on judges' scores) is 0.85. This correlation is 0.9 across all US seasons.

More importantly, however, both models suggest the presences of the J-curve effect, which exist after controlling for competitor abilities *and* interference in the allocation sequence.[5]

Given our data include judges' scores for all contestants and episodes, we now undertake an analysis based directly on these scores. We begin by examining non-parametric local linear and quadratic specifications that relate average judge scores to the (relative) order. Pooling the data over all seasons, Figure 2 clearly displays a J-shape for both the UK and US data.

Model specifications examining 'average judge score' for the UK and US data are set out in Tables 3 and 4, respectively. Models (1) to (4) employ OLS, whereas Models (5) to (8) employ a (within) FE structure. Once again, in addition to the relative (sequence) 'order' variable, a dummy variable for being 'first' in the sequence of dances is included to investigate a possible J-curve effect. Furthermore, quadratic specifications are also considered in this respect. Sets of dummy variables control for season, week (within season), and dance style.[6]

OLS Models (2) and (3) suggest a J-curve effect in both the UK and US data. However, the 'first' (in sequence) effect disappears in FE Models (6) and (7). This could be the result of producers purposefully scheduling a 'strong' couple first in the show to give the episode a strong and positive start. Essentially, once couple FEs are included, there is no 'first' effect.[7] Furthermore, the quadratic models provide no evidence of quadratic effects for the same reason accounting for the disappearing 'first' effect.[8]

That the 'first' effect is not present in the 'average score' models differs with the results presented in the 'bias' models. This relates specifically to the different dependent variables considered. In particular, while the first contestant might generally receive close to the 'average' score (after controlling for contestant fixed effects), they also rarely feature in the bottom positions, which generates the (positive) 'first' effect (again after controlling for contestant fixed effects).

Given the FE model controls for contestants' abilities, the empirical evidence suggests increasing scores from judges' evaluating contestants later in the sequence. What is the source of this bias? As contestants are scored after each dance, the memory-induced bias relating to 'recency' should not factor. Furthermore, as the 'bias' models suggest no allocation bias we do not believe this is producer induced. We are then left with a type of 'grade inflation', which may relate to the euphoria of the contest combined with audience effects.

[5] Page and Page (2010) provide an eloquent synthesis of this process.

[6] In the UK data, 13 dance styles are categorised, and in the US data, 30 styles are categorised.

[7] In these models, we preference FE over RE given the wider set of explanatory variables included and associated rejection of the Hausman test. We note that the parameter estimates from the RE specification are very similar in qualitative terms.

[8] The FE models do not use seasonal effects as the within specification makes these redundant.

4 Summary and concluding remarks

This study examines sequence order bias in a contest format featuring both expert panel judgement and a popular vote. Contrary to expectations from expert panel outcomes assessing on objective performance criteria (applied after each performance), the UK and US data features clear evidence of sequence order bias effects. These effects resemble a J-curve where there is both a 'first' (in sequence) effect and 'order' effect, such that the first and later performing contestants disproportionately gained higher expert panel scores.

Although we believe managed choice of opening performers may play a role, we suggest that the key sequence biases observed can be interpreted as a type of 'grade inflation' in the expert panel's scoring. In particular, the 'order' effect may derive from studio audience pressure akin to the previously cited evidence on spectator pressure on referees.

When popular votes augment the expert panel scores, the 'primacy' and 'recency' biases observed in previous studies appear to reinforce and even amplify the J-curve effect. Our results add to the evidence that creators of such contests should pay careful attention to evaluation design given the various biases that may exist, including designs where they might not be expected.

References

Dowie, J. (1982). Why Spain should win the world cup. *New Scientist*, 94(1309):693–695.

Lenten, L., Crosby, P., and McKenzie, J. (2019). Sentiment and bias in performance evaluation by impartial arbitrators. *Economic Modelling*, 76:128–134.

Mantonakis, A., Rodero, P., Lesschaeve, I., and Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, 20(11):1309–1312.

Nevill, A. M., Balmer, N. J., and Williams, A. M. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, 3(4):261–272.

Page, L. and 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.

Figure 1: Elimination bias vs. sequence order

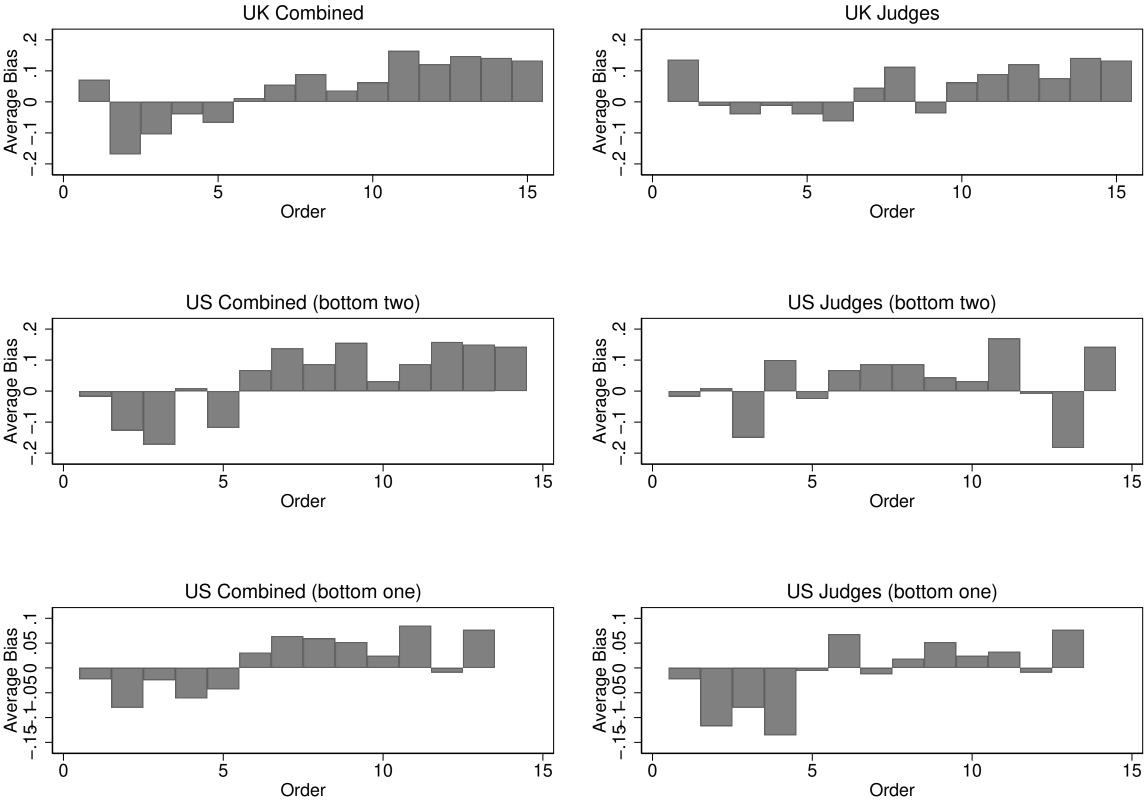

Figure 2: Non-parametric and quadratic models of average score

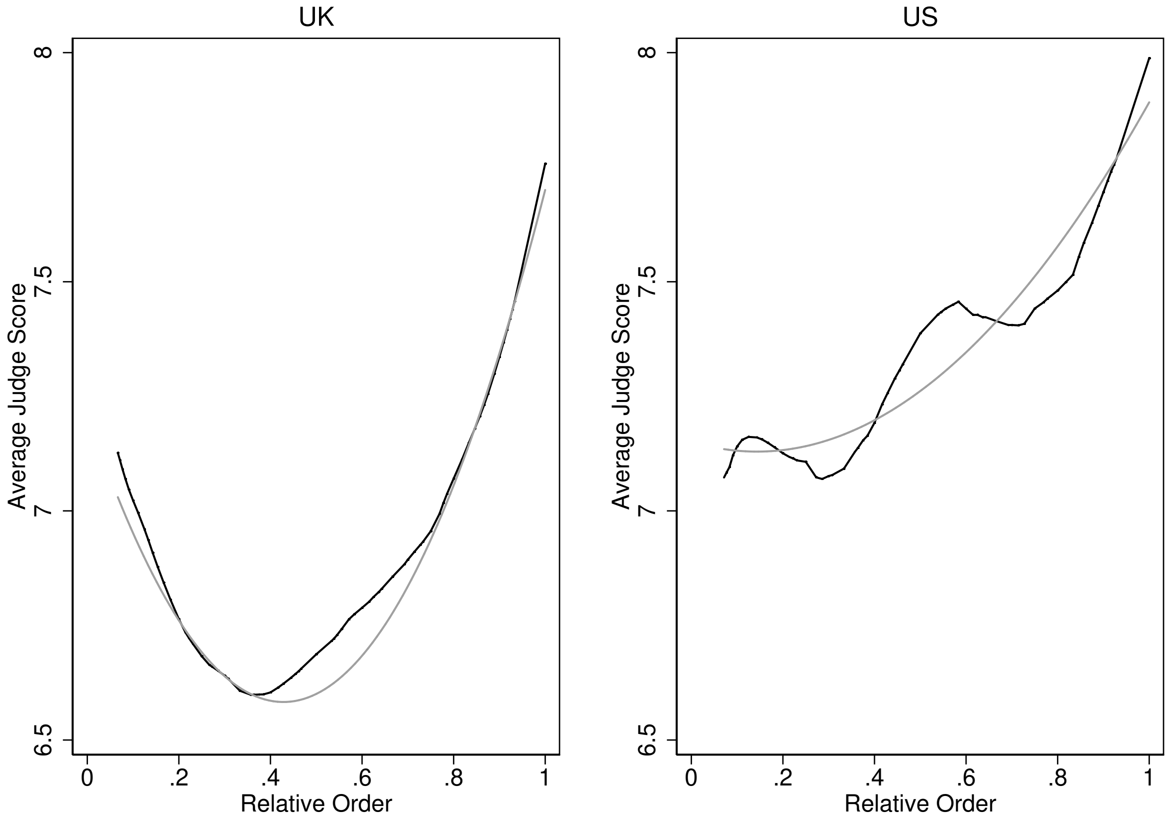

Table 1: UK regressions of bias on sequence

	Combined		Judges	
	FE	RE	FE	RE
Order	0.379***	0.384***	0.188***	0.196***
	(0.005)	(0.004)	(0.008)	(0.007)
First	0.266***	0.269***	0.219***	0.225***
	(0.004)	(0.004)	(0.007)	(0.007)
Const.	-0.237***	-0.240***	-0.106***	-0.112***
	(0.003)	(0.003)	(0.005)	(0.005)
Obs	1045	1045	1045	1045
Groups	199	199	199	199
R^2	0.884		0.545	
Hausman (p-value)	4.45	(0.11)	6.05	(0.05)

Notes: Dependent variable is ‘bias’. ‘Order’ is relative (sequence) order. ‘First’ is dummy for first in sequence. FE and RE are fixed and random effects at couple level, respectively. Week and Dance are sets of dummy fixed effects. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, ***$p < 0.01$.

Table 2: US regressions of bias on sequence

	Bottom Two				Bottom One			
	Combined		Judges		Combined		Judges	
	FE	RE	FE	RE	FE	RE	FE	RE
Order	0.409***	0.424***	0.219***	0.229***	0.186***	0.189***	0.206***	0.223***
	(0.018)	(0.014)	(0.021)	(0.016)	(0.007)	(0.005)	(0.010)	(0.007)
First	0.182***	0.188***	0.050***	0.063***	0.063***	0.069***	0.095***	0.110***
	(0.016)	(0.013)	(0.019)	(0.015)	(0.006)	(0.005)	(0.009)	(0.007)
Const.	-0.248***	-0.257***	-0.105***	-0.115***	-0.108***	-0.111***	-0.145***	-0.156***
	(0.012)	(0.009)	(0.013)	(0.011)	(0.004)	(0.003)	(0.006)	(0.005)
Obs	400	400	400	400	533	533	533	533
Groups	169	169	169	169	188	188	188	188
R^2	0.6836		0.355		0.7027		0.5777	
Hausman (p-value)	1.58	(0.45)	1.38	(0.50)	2.16	(0.34)	9.56	(0.01)

Notes: Dependent variable is ‘bias’. ‘Order’ is relative (sequence) order. ‘First’ is dummy for first in sequence. FE and RE are fixed and random effects at couple level, respectively. Week and Dance are sets of dummy fixed effects. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, ***$p < 0.01$.

Table 3: UK regressions of average (judge) score on sequence

	OLS				Within FE			
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Order	0.837***	1.388***	1.378***	-0.063***	0.582***	0.491***	0.649***	0.508***
	(0.149)	(0.172)	(0.151)	(0.250)	(0.094)	(0.115)	(0.096)	(0.145)
First		0.998***	1.042***			-0.148	0.047	
		(0.162)	(0.146)			(0.106)	(0.091)	
Order sq				0.007***				0.001
				(0.002)				(0.001)
Const.	6.455***	6.049***	5.075***	5.994***	6.595***	6.661***	6.172***	6.220***
	(0.092)	(0.112)	(0.313)	(0.332)	(0.058)	(0.074)	(0.129)	(0.127)
Couple FE	No	No	No	No	Yes	Yes	Yes	Yes
Season	No	No	Yes	Yes	No	No	No	No
Week	No	No	Yes	Yes	No	No	Yes	Yes
Dance	No	No	Yes	Yes	No	No	Yes	Yes
Obs	1276	1276	1276	1276	1276	1276	1276	1276
Groups					205	205	205	205
Adj. R^2	0.024	0.051	0.285	0.265	-0.151	-0.150	0.227	0.228
LL	-2347	-2328	-2130	-2147	-1548	-1547	-1280	-1280

Notes: Dependent variable is 'average judge score'. 'Order' is relative (sequence) order. 'First' is dummy for first in sequence. 'Order sq' is squared relative order. Models (1)-(4) are OLS. Models (5)-(8) are within (couple) fixed effects. Season, Week, and Dance are sets of dummy fixed effects. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, ***$p < 0.01$.

Table 4: US regressions of average (judge) score on sequence

	OLS				Within FE			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Order	0.846***	1.044***	0.962***	0.535**	0.661***	0.605***	0.603***	1.128***
	(0.118)	(0.138)	(0.118)	(0.249)	(0.091)	(0.108)	(0.087)	(0.169)
First		0.364***	0.308***			-0.094	-0.128	
		(0.132)	(0.114)			(0.098)	(0.079)	
Order sq				0.002				-0.004***
				(0.002)				(0.001)
Const.	6.922***	6.776***	6.080***	6.357***	7.024***	7.064***	6.150***	6.036***
	(0.074)	(0.090)	(0.485)	(0.498)	(0.055)	(0.069)	(0.289)	(0.289)
Couple FE	No	No	No	No	Yes	Yes	Yes	Yes
Season	No	No	Yes	Yes	No	No	No	No
Week	No	No	Yes	Yes	No	No	Yes	Yes
Dance	No	No	Yes	Yes	No	No	Yes	Yes
Obs	1349	1349	1349	1349	1349	1349	1349	1349
Groups					317	317	317	317
Adj. R^2	0.036	0.041	0.320	0.317	-0.244	-0.245	0.220	0.224
LL	-2215	-2211	-1948	-1951	-1473	-1472	-1136	-1132

Notes: Dependent variable is ‘average judge score’. ‘Order’ is relative (sequence) order. ‘First’ is dummy for first in sequence. ‘Order sq’ is squared relative order. Models (1)-(4) are OLS. Models (5)-(8) are within (couple) fixed effects. Season, Week, and Dance are sets of dummy fixed effects. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, ***$p < 0.01$.