Last shall be first:
A field study of biases in sequential performance evaluation on the Idol series.

Lionel Page*
Westminster Business School
University of Westminster 35, Marylebone Road
NW1 5LS, London, UK
l.page@wmin.ac.uk

Katie Page
Department of Psychology
Royal Holloway University of London
TW20 0EX, Egham, UK
katie.page@rhul.ac.uk

March 4, 2008

*Corresponding author. Tel: +44 2 07911 5000 (2706)

Abstract

When performances are evaluated they are very often presented in a sequential order. Previous research suggests that the sequential presentation of alternatives may induce systematic biases in the way performances are evaluated. Such a phenomenon has been little studied in Economics. Using a large data set of performance evaluation in the Idol series (N=1522), this paper presents new evidence about the systematic biases in sequential evaluation of performances and the psychological phenomenons at the origin of these biases.

JEL codes: D81, Z1
Keywords: order effects, memory, television show

We frequently make judgments and decisions about information which is presented to us in a sequential manner. This in particular is the case when we have to quickly assess the performance of individuals within a pool of contestants: job interviews, singing auditions, political debates, or even dating evenings.

The psychological literature suggests that sequential presentation of information may influence the way each piece of information is processed and recorded. Studies in economics (Neilson 1998) and marketing (Novemsky and Dhar 2005) have also found that a choice in a situations of sequential choices may be dependent of the history of the sequence. This issue is of special importance for situations of performance evaluation. If there is any effect of the order in which people are assessed on the final evaluation of individual performances, it means that the evaluation process is biased. Stated simply, what should be completely irrelevant information (the passing order) plays a significant role in the evaluation process.

The issue of potential bias in performance evaluation raises two main concerns: efficiency and fairness. First, from the perspective of the assessor, any bias in the evaluation process results in a loss in terms of efficiency because the best options may eventually not be selected. Second, from the perspective of the contestant, any bias in the evaluation process raises the question of the fairness of the selection process: are some contestants disadvantaged relative to others for irrelevant reasons?

If there are biases in evaluation processes involving a sequential ordering of the contestants/options, we need to be aware of these in order to design strategies to minimize their adverse effects or ensure that outcomes are as fair and efficient as possible.

Paradoxically, few studies have attempted to assess empirically the presence of systematic biases in the sequential evaluation of performance (Bruine de Bruin 2005). More specifically, in Economics, the fairness and efficiency of performance evaluation procedures have mostly been studied relative to the possible biases arising from the judges incentives (Prendergast and Topel 1993, Clerides and Stengos 2006) and from discriminating preferences (Goldin and Rouse 2000, Segrest Purkiss, Perrewé, Gillespie, Mayes, and Ferris 2006). The Economic literature has largely ignored the possible distortions arising from the pure cognitive biases in the evaluation of performance. Such biases, if significant and of practical importance, must however been studied carefully in order in order to limit their detrimental effects on the efficiency and fairness of the selection procedures relying on the evaluation of performances.

Using a unique dataset on the Idol series spanning competitions from 8 countries (Australia, Brazil, Canada, Germany, India, Netherlands, United Kingdom, USA), this paper contributes to our understanding of order biases in performance evaluation in a naturalistic setting. Because of their generic format, the Idol shows provide a large set of identical situations where a set of individuals have to perform sequentially and are assessed by television viewers who vote for them.

The statistical analysis of this large dataset of 1,522 performances over 165 shows confirms some of the previous empirical literature on ordering effects and contributes to furthering our understanding of the underlying psychological phenomena of these effects. Our results suggest that systematic biases in sequential evaluation of performance arise through two parallel processes: the effect of the ordering on the propensity to remember each candidate, and the propensity to assess a contestant by comparing him or her to the previous contestant(s).

The remainder of the paper is organized as follows, Section 1 presents the literature on sequential biases in performance evaluation, Section 2 presents our dataset and Section 3 our results. Section 4 concludes.

1 Sequential biases in performance evaluation

There are two main reasons why biases may result from sequential ordering. The first is that judges may not remember equally well the different performances in the sequence, and second, the criteria/benchmark of the evaluation may change over time. For example, the evaluation of a performance may be dependent on the history of previous performance(s).

These potential caveats may produce two types of biases. First, ordering biases may result because your performance evaluation is conditional on your passing order. The second potential bias is that the evaluation of one's performance may directly depend on the quality of the previous performance(s). We will call these two types of biases respectively "sequential order bias" and sequential history bias".

1.1 Sequential order bias

Few studies have addressed the effect of order on judgments of performance. Generally the research evidence indicates that later serial positions benefit from more positive evaluations . The evidence comes from several naturalistic studies on performance in competitions, including a study on international synchronized swimming competitions (Wilson 1977), work on the Queen Elizabeth Contest for violin and piano (Glejser and Heyndels 2001), and studies of the Eurovision song contest (Bruine de Bruin 2005) and ice skating competitions (Bruine de Bruin 2005, Bruine de Bruin 2006).

Wilson (1977) showed that there was a significant negative correlation between serial positions and final ranks in the 1973 World Championship synchronized swimming championships and an amateur meet held in the same year such that better rankings tended to be in later serial positions. Final ranks in each competition were determined by two rounds of performances, each judged by a different experienced jury.

An evaluation of the judgments by 15 experts in the Queen Elizabeth Contest for classical violin and piano by Glejser and Heyndels (2001) showed that musicians who performed on a later day in the competition received better judgments. Moreover, higher overall rankings were also given for performances scheduled later in the week and later in the evening (Glejser and Heyndels 2001).

Bruine de Bruin (2005) examined the effect or order in both the Eurovision song contest and ice skating judgments. She found an increasing linear trend such that contestants who were in the later serial positions had significantly higher ratings than those in the earlier positions. This effect was also found in her follow up study on ice skating (Bruine de Bruin 2006) with a larger data set.

Two potential explanations exist in the literature for this observed order bias. First Bruine de Bruin (2005) explain their results through a direction of comparison effect. Specifically, they posit that as each new option is presented judges search for unique features (positive or negative) in the performance and, if found, these influence upwardly (for positive unique features) and downwardly (for unique negative features) the judgments, because more weight is given to these unique options rather than any overlapping features of the performance. Overall, they conclude that the direction-of-comparison effect is most prominent in tasks that promote sequential judgment, and in options with unique positive features (Bruine de Bruin and Keren 2003).

They further speculate that the direction-of-comparison effect may have contributed to the linear order effects found in jury evaluations of world-level figure skating contests (Bruine de Bruin and Keren 2003), international synchronized swimming competitions (Wilson 1977), the Eurovision Song Contest for popular music (Bruine de Bruin 2005), and the Queen Elizabeth Contest (Glejser and Heyndels 2001). However, this would only be the case if the judges were focused on the unique positive features of each performance, which may or may not have been the case.

A second possible explanation for the empirical results relates to memory. There is a well established literature on the effects of order on memory. The serial position effect is the phenomenon demonstrating that recall accuracy (usually for words) varies as a function of an item's position within a list (Murdock 1962). Specifically, there are two main effects: the primacy and recency effect. When asked to free recall items from a list participants generally remember better those stimuli at both the beginning (primacy effect) and end (recency effect) of a sequence, resulting in a roughly u-shaped curve. The serial position effect is a robust well researched phenomenon in the cognitive psychological literature (Glanzer and Cunitz 1966, Burgess and Hitch 1999, Gershberg and Shimamura 1994). It has also be shown that memory may play a critical role in economic decisions (Devetag and Warglien 2003, Devetag and Warglien 2007).

These serial position effects have been demonstrated both in the laboratory (Singh and Cole 1993, Snyder and Harrison 1997) and in naturalistic settings (Terry 2005, Pieters and Bijmolt 1997). Different memory mechanisms have been proposed to underlie the primacy and recency effects, with primacy effects linked to long term memory and recency effects explained through short term memory mechanisms (Glanzer and Cunitz 1966). Moreover, several factors have been found to influence or alter their effects, for example distinctiveness (Neath and Crowder 1996), emotional content (Rubin and Friendly 1986, Maratos, Allan, and Rugg 2000, Snyder and Harrison 1997), prolonged distraction (Glenberg, Bradley, Stevenson, Kraus, Tkachuk, Gretz, et al. 1980) and the length of the series (Anderson, Bothell, Lebiere, and Matessa 1998). Generally though, holding other factors constant, first and last items are remembered better.

Whilst these memory explanations have been seldom linked to the evaluation of sequential performance extrapolating the results would suggest that contestants who are in earlier and later positions will benefit positively as a result of their performances being better remembered.

1.2 Sequential history bias

The second possible bias in the sequential evaluation of performance is that each person's performance evaluation may depend on the performance of the previous person relative to whom they are often implicitly compared. For each judgment in a given sequence (with the exception of the first judgment), it is the case that the judge has already very recently evaluated another target on that same dimension. Therefore, the knowledge the judge has activated to make that previous judgment is highly accessible at the time the next judgment has to be made. Consequently, this knowledge of the previous judgment is likely to influence the subsequent judgment(Damisch, Mussweiler, and Plessner 2006).Thus, the evaluation of a target at almost any point of the sequence is likely to be affected by the information that was activated during the preceding judgment of another target on that dimension (Damisch, Mussweiler, and Plessner 2006, 167).

Mussweiler, Rüter, and Epstude (2004) selective accessibility model outlines two main comparison processes-contrast and assimilation-that take place during the assessment of two consecutive stimuli. Contrast occurs when judges focus on differences in the stimuli, and assimilation occurs when the focus is on similarities. More precisely, the direction of the influence is determined by the perceived similarity between the two sequential stimuli. A priori it is not clear what phenomenon is likely to be at work in a sequential performance evaluation, but regardless of its nature it is likely to create biases in the individual evaluation of performances because the evaluation of a contestant's performance will be depend on the performances of the previous contestant.

Damisch, Mussweiler, and Plessner (2006) examined sequential performance judgments in both the 2004 Olympic Games and data gathered in a laboratory setting. Their aim was to apply the concepts in the selective accessibility model (Mussweiler, Rüter, and Epstude 2004) to sequential judgments in sport. Their results demonstrated that the score of an athlete increases with increasing scores of his or her immediate predecessor and decreases with decreasing scores of his or her predecessor, showing assimilation rather than contrast. Moreover, this effect carries on after the first person such that the correlation between a target and subsequent targets, whom are not immediately after the target (but second third etc), are also significant. According to research by Mussweiler, Rüter, and Epstude (2004) and Gentner and Markman (1994) unless otherwise instructed judges tend to search for similarities in the performances of people, that is, assimilation often appears to be the default judgmental outcome, resulting in significant positive correlations between performances.

Overall, there seems to be two biases at work in influencing one's overall performance ratings. First, there is the effect of the overall order on performance where either (a) first and last positions are favored (cf. the memory literature) and/ or (b) there is an increasing linear trend. In addition, there is a second effect which involves a direct comparison process, where the outcome of your performance is influenced by the evaluation of your predecessor. In cases where you are evaluated immediately after someone who is judged favourably you are also likely to be judged well and vice versa for an unfavourable performance. However, this is only the case when the process at work is assimilation rather than contrast.

This paper investigates two biases in the sequential evaluation of performance in a large data set from a naturalistic setting. Its unique contribution is two fold. First, no previous work has evaluated these two biases concurrently, therefore this paper adds to the existing work by enabling a direct comparison of these two processes in sequential order effects on performance evaluation. This is extremely important because it will enable us to isolate what factors are contributing to an observed ordering effect in performance and provide clearer theoretical implications.

Second, this paper uses a large, multicultural dataset which has the advantage of ecological validity and generalisbility. A large majority of the previous studies of these order biases tend to be laboratory based or naturalistic studies using much smaller or restricted datasets. Our paper is unique in this respect and hence provides a strong base for testing the theoretical predictions.

2 The data

Our data consist of observations of the ranking of contestants in live shows for several pop Idol series: Australia (Australian Idol: 2003, 2004, 2006, 2007; X-factor: 2005), Brazil (Ídolos Brazil: 2007), Canada (Canadian Idol: 2003, 2004, 2005, 2006, 2007), Germany (Deutschland sucht den Superstar: 2003, 2004, 2006, 2007), India (Indian Idol: 2006, 2007), Netherlands (Idols: 2005; X factor: 2006), UK (X-factor: 2004, 2005, 2006, 2007), and the USA (American Idol: 2002, 2003, 2004, 2005, 2006, 2007). All of these shows share the same format in their final stage, specifically, the final set of contestants (10 to 13 depending on the series) are progressively eliminated one by one after each show. In each session participants have to perform a new song. Their performance is then assessed by television viewers who can vote for their preferred performance. The votes are tallied and one of the last two (or three) contestants who have received the fewest votes from the public is then eliminated (sometimes this last step is determined by a choice from the judges).

The generic format of these shows, which is almost identical across countries and seasons, provides a unique opportunity to study the effects of ordering on the evaluation of individual performance. In addition, the variety of countries in our sample ensures that our results are not idiosyncratic to a given culture or to a given series.

For each season, we observe the final sessions where candidates have to perform one song one after the other, before the public is allowed to vote for them. We do not analyse the very final stage of the competition, when four or five competitors are left and they each sing two or more songs. We therefore observe only sessions where there are between 5 and 13 competitors singing one song and one or two competitors being voted off at the end of each show. This data has been collected on various online sources: wikipedia.org, tv.com and the shows' websites.

Table 1: Breakdown of the number of shows by country and number of contestants

Contestant	AUS	BRA	CAN	GER	IND	NED	UK	USA	Total
5	4	0	5	3	2	0	1	1	16
6	4	1	5	4	2	1	4	5	25
7	4	1	4	4	2	2	4	6	25
8	5	1	5	3	2	2	4	6	26
9	3	1	4	4	2	1	3	4	21
10	4	1	4	4	1	2	2	6	22
11	3	0	1	2	1	1	3	4	14
12	4	1	0	0	1	0	3	4	13
13	0	0	0	2	1	1	0	0	3
Total	31	6	28	26	14	10	24	36	165

Due to the marketing policy of the show, and in order to maintain the highest suspense during the competition the shows do not reveal the exact proportion of votes for each contestant (with the exception of the German competition in the last three years). However, we do have some information about the rankings of the contestants because the bottom two, three or four competitors are revealed each time.

3 Method

To assess the existence of a bias in the evaluation of contestants' performance, we will compare compare the empirical probability to be "safe" during one show to the theoretical probability if there is no biases from the sequential ordering from the contestant.

Imagine a series of shows with a constant number N of contestants and suppose that these contestants have the same qualities (hence the same a priori probability to be safe). Let $b_k \in \{2,4\}$ be the number of individuals in the bottom tier for a show k, the probability to be safe for a contestant is:

$$p_k = 1 - \frac{b_k}{N}$$

Suppose now that the ordering of the performances in the live show has an impact on the evaluation of the performance by the television viewers. Some participants will be favored by their position in the series and other disadvantaged. Lets call $bias(X,Z)$ this systematic departure from the theoretical probability of being safe where X is a set of variable characterising the position of the contestant in the passing order, and Z a set of variables describing the characteristics of previous contestants. The probability to be safe for a participant in this position is

$$p_i = 1 - \frac{b_k}{N} + bias(X,Z)$$

Suppose that, in this simple situation, we want to estimate the bias linked with every position i of the order, $E(bias(i,Z)|i)$, we could compare the theoretical probability to be safe $p_T = 1 - b_k/N$ to the actual frequency of safe contestants in each position i, $\widehat{p_i} = \sum \mathbb{1}_{\{i \text{ is safe}\}}/N_s$, where N_s is the number of shows observed:

$$E(bias(i,Z)|i) = \sum \frac{\mathbb{1}_{\{i \text{ is safe}\}}}{N_s} - \frac{b_k}{N}$$

Our data is slightly more complex than this example since the number of contestants varies across the shows. To estimate $E(bias(X,Z)|X,Z)$, we build the variable $bias_{jk}$, which, for a participant j performing in the show k takes the value:

$$bias_{jk} = \mathbb{1}_{\{j \text{ is safe}\}} - \left(1 - \frac{b_k}{N_k}\right)$$

By definition, we have $E(bias(X,Z)|X,Z) = E(bias_{jk}|X,Z)$. We can then define the two biases found in the literature as:

Definition 1 (Sequential order bias) *There is a sequential order bias as soon as for any variable x_j characterising the position of a performance j in the passing order:*

$$E(bias_{jk}|x_j) \neq 0$$

Definition 2 (Sequential history bias) *There is a sequential history bias as soon as for any variable z characterising the previous candidates:*

$$E(bias_{jk}|z) \neq 0$$

The following sections will consecutively study these two possible biases.

4 Sequential order bias

A sequential order bias arises when a candidate is advantaged or disadvantaged for his/her position in the order. To study this possible bias, we will first look at the the value of $E(bias_{jk}|i)$ which represents, for a given position in the order of appearance i, the difference in percentage points between the actual and theoretical probability to be safe. It therefore measures the advantage/disadvantage the position confers to a contestant in terms of the probability to be safe. Specifically, if $E(bias_{jk}|i)$ is positive then a contestant j in position i is more likely to be safe, and if $E(bias_{jk}|i)$ is negative he/she is less likely to be safe.

*** Figure 1: Bias in performance evaluation by position order ***

Figure 1 presents the mean bias per order over the whole set of orders. A clear pattern emerges which shows a positive trend as the order increases. However, this graph is imperfect because the relative position of each order may be different. For example the 5th order will be the last one in some situations, while in other situations it will be located in the middle between the beginning and the end of the series. In this graph the last order also consists of different orders, for example sometimes it is 5th, 9th or 11th. Figures 2 and 3 present the decomposition of the ordering effect for the sessions which have between 5 and 12 contestants. The last contestants appear to benefit from a positive bias, while contestants in the middle of the order (especially closer to the beginning) seem to be disadvantaged.

*** Figure 2: Order effect for each type of session ***
*** Figure 3: Order effect for each type of session ***

In order to summarize the effects at the beginning and at the end of the order, Figure 4 compares the evolution from the beginning of the order to the evolution when looking at the reverse order. The last contestants appear to have a significant advantage relative to the contestants in other positions.

*** Figure 4: Bias in performance evaluation at the beginning and the end of the series ***

Overall, these results suggest that there seems to be an increasing linear trend such that contestants in the later positions have an advantage relative to those contestants in earlier positions. The worst positions in terms of bias seem to be positions two and three.

One potential caveat of the research concerns the allocation process of the contestants. The above analysis assume the random ordering of contestants to positions. What if this is not the case? In fact, there are two main reasons to think that the ordering is not random.

First, the goal of the production is to maximise the entertainment level and if there is not a strict rule about the random allocation of contestants, this could produce a spurious correlation between ordering and results. For instance, better quality contestants could be more likely to be placed in some specific positions (like the beginning or the end) just for production purposes. This implies that even if there were no ordering effect at all, a selection bias could induce some differences between the probability of success of different positions.

Second, the production could have an agenda regarding the applicants and be willing to keep good contestants longer (because they will attract more viewers later on for instance). So, if there is any ordering effect, they could use it to advantage/disadvantage some contestants. This implies that if there is an ordering effect, the magnitude of this effect could be biased by a selection effect. In order to control for this potential caveat, we implement fixed effect models and estimate the ordering effects while controlling for the ability of the contestant.

To analyse the effect of the ordering on the evaluation of the performance of contestants, it is possible to use a linear regression model with the variable *bias* as a dependent variable. Given that contestants in general perform more than once in the shows, we have repeated observations for contestants, and as arguably contestants vary in quality, the OLS estimator is not efficient and a random effect model must be used instead. The random effects model relies on the same identification assumption than an OLS model: the allocation of the order numbers is random. If, on the contrary, the allocation of the order numbers is not random there is a risk of selection bias in the sense that different positions in the order may be more or less systematically allocated to contestants with differing levels of ability. We must then use a fixed effect model which is a *within* estimator. It estimates, for a given contestant, what is the effect to have a given position or another in the show. Therefore, we estimate the following model:

$$bias_{jk} = \beta_0 + X_{jk}\beta + u_j + \varepsilon_{jk} \tag{1}$$

where X_{jk} is a vector of variables relative to the order i of the participant j in the show k. If the allocation of the order numbers is random and if there is no order effect, no variable x from X_{jk} should have a significant coefficient. The term u_j is an individual effect specific to the individual j. Given that the result for each contestant is not independent of the result of other contestants within a given session, these models are estimated with a clusterised robust variance matrix with the sessions as clusters.

For all sessions the order variable was normalised between to 0 (first) and 1 (last). A dummy variable was created to capture the difference between being the first to perform (1) and all other positions (0). Table 2 presents the regression results. The first three columns are random effect estimations, they are more efficient and well identified if the ordering of candidates is not linked with their specific characteristics. The last three columns are fixed effect estimations, they are unbiased even if the ordering of contestants depends on their specific characteristics.

Table 2: Regression: the ordering effect on performance evaluation

	Dependent variable: bias			
	Random effects		Fixed effects	
	(1)	(2)	(3)	(4)
Order	0.202***	0.265***	0.181***	0.234***
	(6.25)	(6.67)	(5.07)	(5.70)
First		0.111*		0.092
		(2.39)		(1.87)
Cons	-0.139***	-0.182***	-0.090***	-0.128***
	(-8.69)	(-7.85)	(-4.58)	(-5.09)
R^2			0.022	0.026
N	1522	1522	1522	1522
Number of group	352	352	352	352
Hausman test p-value			.263	.492

* p<0.05, ** p<0.01, *** p<0.001

Overall the order effect is very significant and implies that, with the exception of the first position, moving one position closer to the end of the show provides an additional 5 percentage point chance of being safe for a contestant. Therefore, ordering plays a major role in the competition, at least to discriminate between contestants close in ability (which is often the case in the latter rounds of such competitions).

The difference between the random effects and fixed effects model gives an indication about the existence of a selection bias of contestants for each position. The coefficients are very close indicating that the order effect is very unlikely to be driven by a selection bias. To test for a significant difference between the coefficients of the two types of model, we need to implement a generalised version of the Hausman test given that we use a matrix of variance robust to the clusterisation of data in our estimation of both models (Wooldridge 2001, p. 291). In both case this test indicates no significant difference in coefficient between the two models (p-values in the last row of Table 2). This result suggests that the random effects models are consistent and must considered as the best estimation procedure available. Practically, this means that there is no reason to think that the results are driven by a non random allocation of the candidates.

Figure 5 presents the estimation of the parametric prediction from the fixed effect model and a non parametric estimation using a local linear regression for greater flexibility. The two curves match very well and this confirms the good calibration of the linear models. The results for the effect of ordering on performance evaluation show a J-shaped curve rather than a U-shaped curve indicating both primacy and recency effects, with a stronger recency effect.

*** Figure 5: Effect of the relative order on performance evaluation ***

5 Sequential history bias

Another bias possibly arising from the sequential ordering of contestants is that the evaluation of a contestant's performance may be influenced by the performance of the previous contestant to whom they may be compared. If there is an assimilation process, we would expect that contestants performing just after a good contestant are more likely to be highly evaluated and to be in safe. On the contrary, if there is a contrast effect, we would expect it to be an disadvantage to perform after a good contestant as this is likely to negatively affect the evaluation of the contestant's performance.

It is possible to have an indicator of the quality of the contestant with the previous results of each contestant. We build the indicator *strong* which is a binary variable indicating if the candidate has always been safe in the previous shows. While lots of contestants are in the bottom only once, when they are eliminated, lots of contestants are in the bottom several times before being eliminated. For each show following the first one, there are two categories of contestant: those who have always been safe before and those who have been in the bottom tier in a previous show. Arguably, for a given show, a contestant who has never been in the bottom tier previously is less likely to be in the lower range of the ranking than contestants who have been in the bottom tier.

Using the variable *strong*, we look at the effect of being preceded by *strong* contestants on the probability to be safe. We therefore estimate the model:

$$bias_{jk} = \beta_0 + X_{jk}\beta + \sum_{h=1}^{6} strong_{i-h} + u_j + \varepsilon_{jk} \tag{2}$$

Where $strong_{i-h}$ is the dummy variable indicating if the contestant who passed h position before have always been safe in previous shows.

Table 3 displays the results of this model. The estimation of the random effect model does not indicate any effect of the quality of previous contestants. However the fixed effects model suggests a strong effect of the previous contestant. The Hausman test indicates that the coefficients in the fixed effects model are significantly different from the coefficients in the random effects model. This suggests that the random effects model is inconsistent. This may be the case if for instance the productions of the shows tend to prevent to have two weak candidate consecutively. The effect estimated in the fixed effects model is then underestimated in the random effects model.

The results of the fixed effects model suggests a significant and important effect of the previous contestant quality on the evaluation of the current contestant performance. When the previous contestant has never been once in the bottom tier before, the current contestant has 10 percentage point more chance to be safe. The coefficients for other previous contestant is also negative but lower and not almost always non significant.

Table 3: Regression: the comparison effect relative to the previous contestant

	Dependent variable: bias							
	Random effects				Fixed effects			
	(1)	(2)						
Order	0.272***	0.288***	0.291***	0.310***	0.251***	0.249***	0.234***	0.239**
	(6.88)	(6.06)	(5.20)	(4.61)	(5.91)	(4.96)	(3.93)	(2.99)
$strong_{i-1}$	0.047	0.043	0.047	0.027	0.108***	0.102**	0.092**	0.056
	(1.84)	(1.51)	(1.53)	(0.82)	(3.90)	(3.21)	(2.61)	(1.47)
$strong_{i-2}$		-0.008	-0.015	0.003		0.034	0.016	0.028
		(-0.30)	(-0.49)	(0.09)		(1.08)	(0.48)	(0.70)
$strong_{i-3}$			0.026	0.014			0.069*	0.062
			(0.84)	(0.41)			(2.13)	(1.58)
$strong_{i-4}$				-0.033				-0.012
				(-0.97)				(-0.30)
Cons	-0.225***	-0.222***	-0.241***	-0.209**	-0.219***	-0.239***	-0.260***	-0.229*
	(-7.17)	(-4.92)	(-3.94)	(-2.65)	(-6.50)	(-5.24)	(-4.04)	(-2.56)
R^2					0.047	0.039	0.033	0.023
N	1339	1156	973	790	1339	1156	973	790
Nb of group								
Hausman p-value					0.001 <	0.001 <	0.001 <	0.001 <

* p<0.05, ** p<0.01, *** p<0.001

6 Test of the random allocation of the contestant

In the previous development we have been careful to control for a possible non random allocation of the contestants in the show. The information on the performances of the contestant on previous shows gives us a way to test more directly for their random allocation during the show. We can test if "strong" contestants who have never been in the bottom tier in previous shows are more likely to be at the end or the beginning of the show, and we can test if there is negative autocorrelation in the allocation of the contestants (weak contestants being more likely to be followed by a strong contestant than by a weak one). To do so, we studied the probability that a contestant at a given order is strong depending on the order and on the quality of the previous contestant:

$$strong_{ik} = \beta_0 + X_{ik}\beta + \sum_{h=1}^{6} strong_{i-h} + \nu_k + \varepsilon_{ik} \tag{3}$$

Where ν_k is fixed effect specific to the show k. This fixed effect approach is necessary as the proportion of candidates having been placed in the bottom tier before may change from one show to the other, typically it can increase with the number of shows in the competition[1]. Assuming that the term $varepsilon_{ik}$ represents an error with a logit distribution, this model is a conditionnal logit. Table 4 present the results of the estimation of this model.

[1] Note that this doe not bias the estimations presented in Table 3

Table 4: Test of the random allocation of the contestants

	(1)	(2)	(3)	(4)
Order	-0.0299	-0.0781	-0.103	-0.367
	(0.22)	(0.25)	(0.37)	(0.63)
First	-0.0939			
	(0.22)			
$strong_{i-1}$		-0.734***	-1.030***	-1.653***
		(0.14)	(0.21)	(0.29)
$strong_{i-2}$			-0.949***	-1.435***
			(0.17)	(0.25)
$strong_{i-3}$				-1.333***
				(0.25)
Observations	1153	987	795	611
R-squared	<0.001	0.02	0.07	0.16

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p¡<.1

These results confirm what our previous analyses suggested. First, there is not systematic bias of the allocation of contestant relative to the passing order: better contestants are not more likely to be toward the end, or at the beginning of the order. Second, there is a negative autocorrelation in the allocation of contestants. Typically, the producers avoid to have two weak contestants in a row.

7 Discussion

Our results suggest that the two mechanisms, memory and direct comparison, both play a role in the order bias. With respect to memory it appears that both primacy and recency effects are implicated when sequentially evaluating performance. Irrespective of ability, contestants who perform first are more likely to be positively evaluated than those who come in second and third positions, which provides evidence of a primacy effect. Contestants who perform in the later serial positions (particularly last position) have the largest advantage with respect to positive evaluations, implying a strong recency effect. The curve showing performance evaluation by serial positions is J-shaped for this dataset implying a much stronger recency effect. These results are partially consistent with those of Bruine de Bruin (2005) who found an increasing linear trend. However, there is divergence with respect to a primacy effect. We find evidence of a small primacy effect while Bruine de Bruin (2005) found no benefit to being in first position. This seems to indicate that memory does play a role in the sequential evaluation of performance.

The second bias we demonstrate is a direct comparison effect with the previous contestant. Specifically, one's performance evaluation is influenced by the evaluation of the previous contestant. If you perform after a weak contestant there is a bias such that you are more likely to be evaluated poorly than if you perform after a strong contestant. Therefore, we find evidence for an assimilation effect with respect to sequential judgments. These findings lend further support to the selective accessibility model of Mussweiler, Rüter, and Epstude (2004). Specifically, our results indicate that judges tend to assess performances based on similarities with the previous contestant and not differences. This is also concurs with evidence from Damisch, Mussweiler, and Plessner (2006). Overall, we show that these two effects both operate and are important explanatory mechanisms in the evaluation of sequential performance.

One factor which could influence these findings concerns the changing performance as a result of being privy to the performance of others. Specifically, it could be plausible that people change their performance (increase level of effort, motivation) after having witnessed the previous performance(s). This mechanism could work in one of two ways. If the task is novel the contestants could learn from the previous performances. However, this is not the case in most tasks which have been studied in the literature (sport and singing competitions) as the task is known in advance. Second, previous performances could act as a benchmark or goal that the future contestant can aim for. Exactly how this process works is unclear and not easy to predict. It could however be an explanation for the dominance of assimilation over contrast because the actual performance is changing rather than the criteria of the judges. One way test this idea would be to investigate these biases in cases where performances are not seen by the contestants, for example in job interviews or private auditions and compare these effects to those cases where the performances are able to be witnessed.

A limitation of the current study is that we do not have information about the number of people who are watching the shows throughout the broadcasts. It is possible, although unlikely in our opinion, that more people are watching the show toward the end of the program and these very same people who miss the beginning of the show also decide to vote. First, it seems likely that the people who are voting are the more ardent fanatics and are less likely to miss the beginning of the show. Second, even if there was a large enough proportion of people voting who miss the early performance(s) then this would mean that we should just see an increasing monotonic trend (assuming people do not vote for people they do not see). Having found a significant primacy effect this result is contrary to this prediction. If anything, these "late voters" would bias downwards the primacy effect which means our estimate of the initial memory effect is likely to be conservative.

Relatively speaking the magnitude of the effect is quite large and therefore is likely to have a significant impact on both the contestants and the judges. Specifically, it is significant enough to raise questions about the fairness of the process from the contestants' perspective and to pose problems in relation to the efficiency of the process from the perspective of the judges. These findings have implications for the way in which performances should be evaluated. At the very least judges (and perhaps contestants) could be made aware of these effects. What they do with this information and how best they assimilate it into their judgments (performances) remains to be studied.

This work also suggests that future research is definitely needed in this area to study in depth these effects. For example, questions that need to be addressed include which is the stronger of these two mechanisms? Do these biases depend of the type of competition and the delay before judging? Also, does making people aware of these biases eliminate them? Moreover, future work needs to study the conditions under which assimilation and contrast are likely to occur in the evaluation of sequential performance. Are certain types of performances (those that are judged on a tight set of criteria) more likely to lead to assimilation effects?

References

Anderson, J., D. Bothell, C. Lebiere, and M. Matessa (1998): “An Integrated Theory of List Memory,” *Journal of Memory and Language*, 38(4), 341–380.

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

——— (2006): “Save the last dance II: Unwanted serial position effects in figure skating judgments,” *Acta Psychologica*, 123(3), 299–311.

Bruine de Bruin, W., and G. Keren (2003): “Order effects in sequentially judged options due to the direction of comparison,” *Organizational Behavior and Human Decision Processes*, 92(1-2), 91–101.

Burgess, N., and G. Hitch (1999): “Memory for serial order: A network model of the phonological loop and its timing,” *Psychological review*, 106(3), 551–581.

Clerides, S., and T. Stengos (2006): “Love Thy Neighbour, Love Thy Kin: Strategy and Bias in the Eurovision Song Contest,” *Centre for Economic Policy Research*.

Damisch, L., T. Mussweiler, and H. Plessner (2006): “Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments,” *Journal of Experimental Psychology Applied*, 12(3), 166.

Devetag, G., and M. Warglien (2003): “Games and phone numbers: Do short-term memory bounds affect strategic behavior?,” *Journal of Economic Psychology*, 24(2), 189–202.

——— (2007): “Playing the wrong game: An experimental analysis of relational complexity and strategic misrepresentation,” *Games and Economic Behavior*.

Gentner, D., and A. Markman (1994): "Structural alignment in comparison: No difference without similarity," *Psychological Science*, 5(3), 152–158.

Gershberg, F., and A. Shimamura (1994): "Serial position effects in implicit and explicit tests of memory," *Learning, Memory*, 20(6), 1370–1378.

Glanzer, M., and A. Cunitz (1966): "Two storage mechanisms in free recall," *Journal of Verbal Learning and Verbal Behavior*, 5(35), 1–360.

Glejser, H., and B. Heyndels (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.

Glenberg, A., M. Bradley, J. Stevenson, T. Kraus, M. Tkachuk, A. Gretz, et al. (1980): "A two-process account of long-term serial position effects," *Journal of Experimental Psychology: Human Learning and Memory*, 6(4).

Goldin, C., and C. Rouse (2000): "Orchestrating Impartiality: The Impact of" Blind" Auditions on Female Musicians," *The American Economic Review*, 90(4), 715–741.

Maratos, E., K. Allan, and M. Rugg (2000): "Recognition memory for emotionally negative and neutral words: an ERP study," *Neuropsychologia*, 38(11), 1452–1465.

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

Mussweiler, T., K. Rüter, and K. Epstude (2004): "The ups and downs of social comparison: Mechanisms of assimilation and contrast," *Journal of Personality and Social Psychology*, 87(6), 832–844.

Neath, I., and R. Crowder (1996): "Distinctiveness and very short-term serial position effects," *Memory*, 4(3), 1–18.

Neilson, W. (1998): "Reference Wealth Effects in Sequential Choice," *Journal of Risk and Uncertainty*, 17(1), 27–48.

Novemsky, N., and R. Dhar (2005): "Goal Fulfillment and Goal Targets in Sequential Choice," *Journal of Consumer Research*, 32(3), 396–404.

Pieters, R., and T. Bijmolt (1997): "Consumer Memory for Television Advertising: A Field Study of Duration, Serial Position, and Competition Effects," *Journal of Consumer Research*, 23(4), 362.

Prendergast, C., and R. Topel (1993): "Discretion and bias in performance evaluation," *European Economic Review*, 37(2-3), 355–65.

Rubin, D. C., and M. Friendly (1986): "Predicting which words get recalled: measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns," *Memory and cognition*, 14(1), 79–94.

Segrest Purkiss, S., P. Perrewé, T. Gillespie, B. Mayes, and G. Ferris (2006): "Implicit sources of bias in employment interview judgments and decisions," *Organizational Behavior and Human Decision Processes*, 101(2), 152–167.

Singh, S., and C. Cole (1993): "The Effects of Length, Content, and Repetition on Television Commercial Effectiveness," *Journal of Marketing Research*, 30(1), 91–104.

Snyder, K., and D. Harrison (1997): "The affective auditory verbal learning test," *Archives of Clinical Neuropsychology*, 12(5), 477–482.

Terry, W. (2005): "Serial Position Effects in Recall of Television Commercials," *The Journal of General Psychology*, 132(2), 151–164.

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

Wooldridge, J. (2001): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Figures

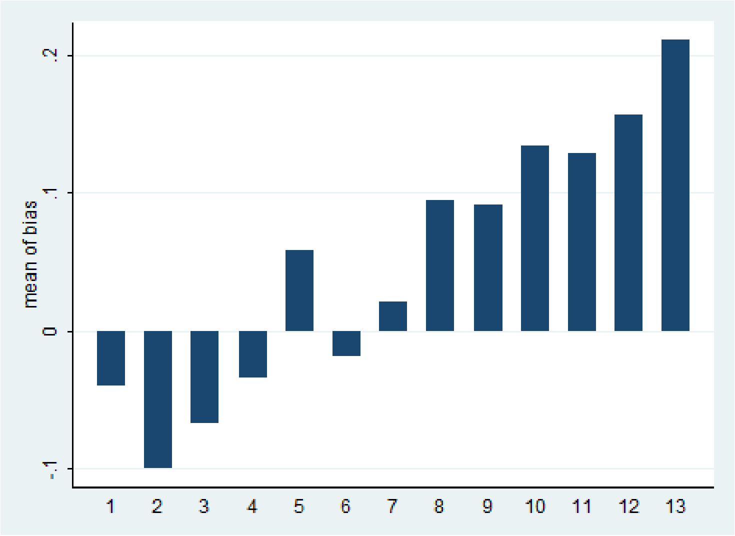

Figure 1: Bias in performance evaluation by position order

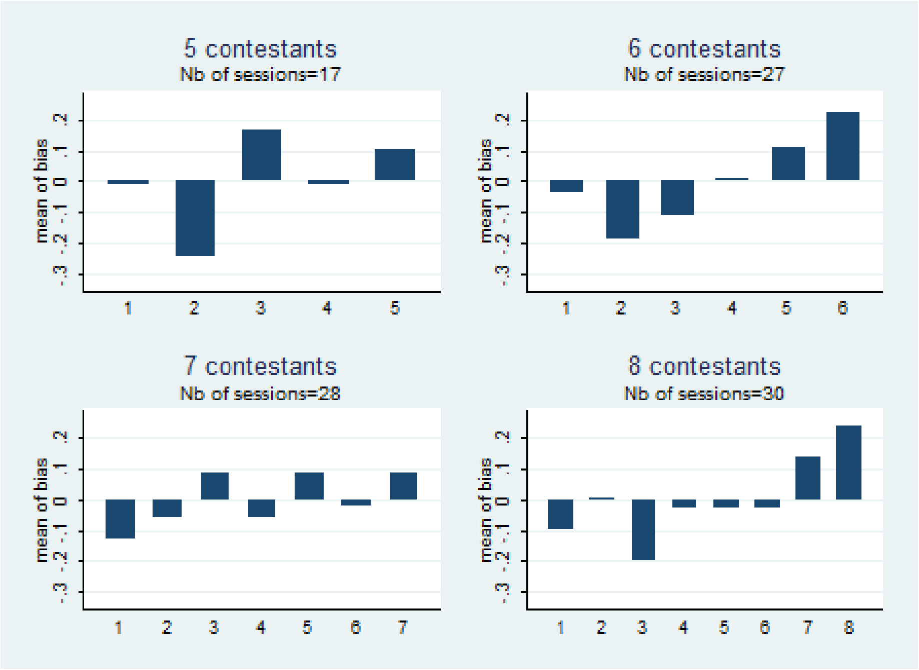

Figure 2: Order effect for each type of session

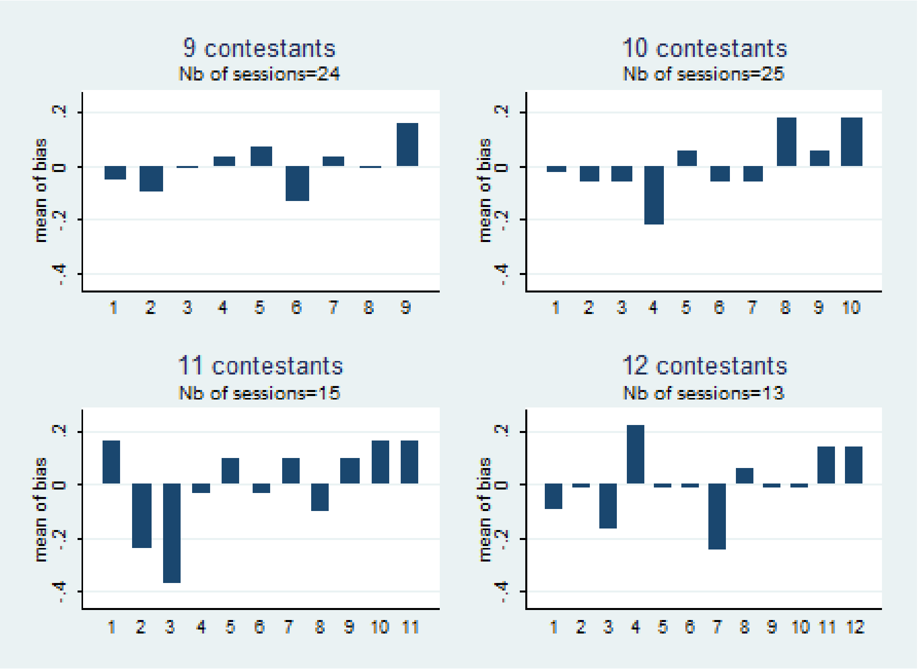

Figure 3: Order effect for each type of session

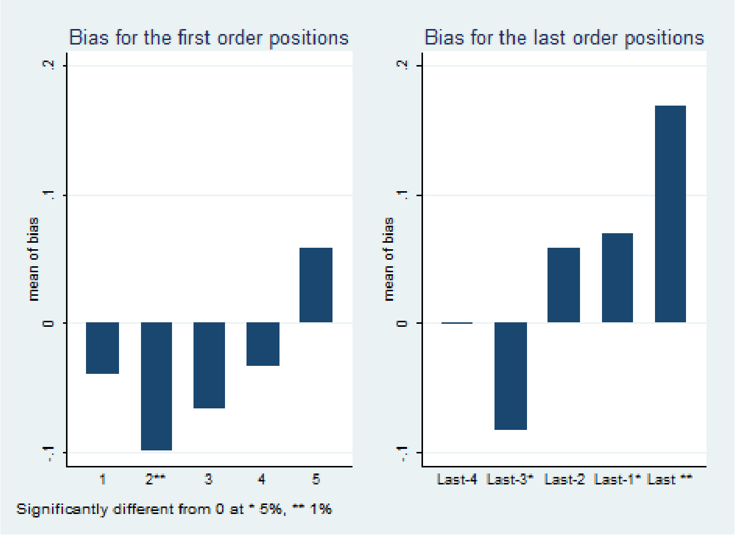

Figure 4: Bias in performance evaluation at the beginning and the end of the series

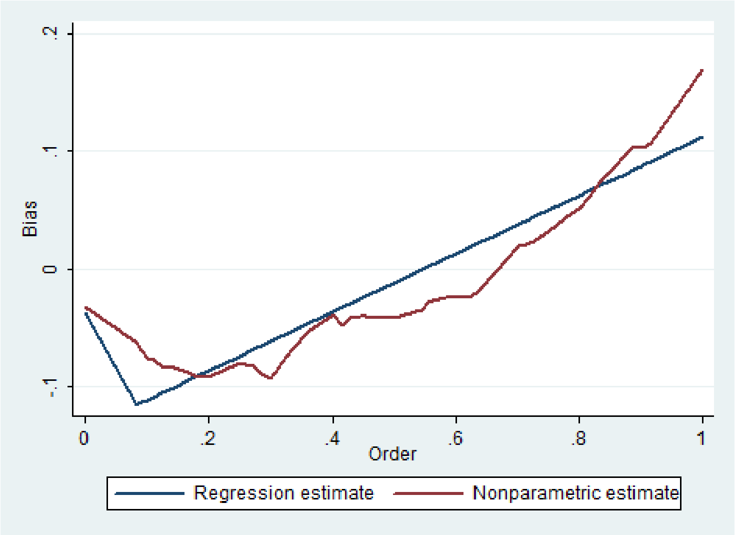

Figure 5: Effect of the relative order on performance evaluation