TESTING FOR BIAS IN ORDER ASSIGNMENT[*]

Darren Grant
Department of Economics and International Business
Sam Houston State University
Huntsville, TX

dgrant@shsu.edu

Abstract: Many real-life situations require a set of items to be repeatedly placed in a random sequence. In such circumstances, it is often desirable to test whether such randomization indeed obtains, yet this problem has received very limited attention in the literature. This paper articulates the key features of this problem and presents three "untargeted" tests that require no a priori information from the analyst. These methods are used to analyze the order in which lottery numbers are drawn in Powerball, the order in which contestants perform on *American Idol*, and the order of candidates on primary election ballots in Texas and West Virginia. In this last application, multiple deviations from full randomization are detected, with potentially serious political and legal consequences.

Keywords: ordering; sequencing; hypothesis testing; randomization; ballot order

JEL Codes: C12, D91

[*] I appreciate the assistance of Elsy Orellana and Jackie Sanchez in assembling data and the contributions throughout from Sheridan Grant, who scraped data, assisted with some of the estimation, and provided constructive commentary and feedback. In addition, a couple of West Virginia county clerks provided valuable insights. This project was supported by grants from Sam Houston State University's Internal Grant Program and College of Business Administration.

Individuals are forced to prioritize in many aspects of life: in judging athletic contests, voting in elections, evaluating applicants for a scholarship or a promotion, and so on. When the rankings thus produced affect others, as in these examples, fairness and efficiency often dictate that we minimize any bias involved in their formation.

One such bias involves the sequence in which the individuals are presented for consideration, which will be called an *ordering*. For example, gymnastics competitions present the contestants one after another; elections present candidates for office in a vertical order on a ballot; final admissions decisions at most elite universities are made one applicant at a time.

In such situations, cognitive bias can affect the way that contestants are assessed, favoring those presented toward the beginning or end of the sequence. A multitude of explanations for such effects have been offered. Krosnick (1991) and Miller and Krosnick (1998) hypothesize that people "tend to evaluate objects with a confirmatory bias," looking for reasons to accept rather than reject them. As they work through the list of options, mental fatigue can lend an advantage to the objects listed earlier in the sequence. Alternatively, many authors, including Mussweiler (2003) and Damisch, Mussweiler, and Plessner (2006), emphasize that the assessment of any one item in a sequence may be influenced by a comparison with its predecessor, in ways that can advantage the later items in the sequence. Salant (2011) argues that order effects are an "inevitable" consequence of "bounded rationality."

They are certainly ubiquitous in real life. A significant body of empirical evidence has arisen in three areas: contests, consumer choice, and voting. In figure skating and music competitions, the evaluations given to contestants tend to increase in sequential order, so that the first contestant tends to be judged most harshly and the last-considered judged most favorably (Bruine de Bluin, 2006; Glejser and Heyndels, 2001; Haan, Dijkstra, and Dijkstra, 2005; Page and Page, 2010; Antipov and

Pokryshevskaya, 2017). On the other hand, in consumer choice, the first good offered is generally preferred (Dean, 1980; Mantonakis et al., 2009; Carney and Banaji, 2012; Novemsky and Dhar, 2005). This primacy effect occurs even more strongly in voting, where over a dozen studies find that the candidate listed first on the ballot receives a boost of 1-5% of the total vote, depending on the contest examined, a phenomenon called the "ballot order effect" (Miller and Krosnick, 1998; Meredith and Salant, 2013; Grant, 2017, and cites therein).

In circumstances like these, there is an incentive for the ordering of contestants to be *manipulated* to advantage preferred individuals or discriminate against ill-favored individuals. Randomly determining the order of contestants minimizes any such favoritism, and many entities have rules to this effect. For example, the order of countries in each year's *Eurovision* song contest is determined by lot, while several states require the order of candidates on each county's primary election ballot to be randomly determined.

In these two examples, orderings are repeatedly conducted for the same set of countries or candidates. In such situations one can test, statistically, whether they have indeed been conducted randomly. If the set of orderings is unlikely to have occurred by random chance, one can conclude that some orderings in the set have been manipulated to favor or disfavor certain contestants. Such manipulation could occur even when prohibited, as the result of a principal-agent problem. For example, Grant (2017) found statistically significant deviations from randomness in multiple primary elections in Texas–a sign that some county election officials were ignoring the law.

To date, however, empirical work has used improvised, ad hoc methods to conduct such a test. Page and Page (2010), examining the program *American Idol*, investigate whether "strong" contestants who scored well in prior rounds are more likely to be presented at the beginning of the show or the end. Meredith and Salant's (2013) analysis of the order of candidates on election ballots uses a similar approach, testing to see if incumbents' ballot positions are evenly distributed. Both of these approaches use limited information, as they ignore the positions of other contestants. Other approaches, though similarly ad hoc, are more inclusive. Ho and Imai (2008) examined the average difference in rank between pairs of letters in randomized alphabets used for ballot orderings in California, while Grant (2017) applied the Fisher Exact Test (later shown to be anti-conservative) to the cross tabulation of ballot positions of candidates in primary elections. Nonetheless, these studies also use limited information, as they rely only on the aggregated tabulation of ranks across letters or candidates, not the individual orderings themselves.

There is thus a need for the systematic development of best-practice methods to test a set of orderings for randomness. Such methods would have greater power to detect deviations from randomness in a variety of circumstances and greater credibility in situations in which failure to randomize appropriately is penalized.

An initial attempt at doing so is a recent paper by Grant, Perlman, and Grant (2020, hereafter GPG). This paper derives two testing methods, described below, that look for deviations from randomness that are specified by the analyst in advance. These *targeted* tests require the analyst to anticipate how any such orderings would be manipulated. In some circumstances this assumption can be reasonable, such as ordering a set of candidates on a ballot, one of which is far more popular than the others. But in most circumstances this is not so. Then these targeted tests are unsuitable. More general, *untargeted* tests are needed that can detect any type of deviation from randomness, statistical power permitting. These tests would be defensible in any context, especially those in which targeted tests are impractical, filling a major gap in the literature.

This paper introduces three such tests and examines their usefulness using both simulations and real-life applications. In suitably large samples, these tests reliably detect systematic manipulation of the orderings with no a priori information required. Such manipulation is observed in one of our applications, with potentially serious legal and political ramifications.

That is this paper's immediate purpose. But it has a larger purpose as well. Applied work in this area has many facets: computational issues, the structure of the data, power considerations, the motivations for deviating from randomness and the nature of the preferences thereby indulged. As this literature is embryonic, most of these have not been previously examined or even articulated as issues; the others have been examined by GPG in a limited fashion suited to the targeted tests developed in that paper. Here we examine them all in depth, in the process delineating the landscape of practical issues empirical analysts can face in testing a set of orderings for randomness.

Accordingly, we analyze an eclectic mix of real-life applications: the sequence in which contestants perform on the popular television show *American Idol*; the order in which the "white balls" are drawn in the largest lottery in the U.S., Powerball; and the order in which candidates are listed on election ballots in Texas and West Virginia. These applications have varied types and strengths of preferences; varied data structures and statistical power; mechanical, pro-social, and anti-social motivations for manipulating the orderings; and wide-ranging computational demands.

This variety drives this paper's ultimate conclusion: there is no single "best test" for the randomness of a set of orderings. The optimal test depends on the circumstances, and can include a test that was developed previously, any one of the tests put forward in this paper, or an adaptation of one of these tests. Researchers are fully equipped for empirical work in this area only after they understand this landscape, for then they can choose the analytical method that best suits the purpose.

I. Three New Tests–and Three Old Ones.

Setup. Following GPG, let there be K *items*, indexed by k = 1...K, which are ordered by each of N *agents*, indexed by i = 1..N. (Alternatively, a given agent could order these items N times.) An ordering is an arrangement or permutation $\Pi \equiv (\pi_1, \pi_2 \dots \pi_K)$ of these items such that item k is assigned *position*, or rank, π_k. These orderings can obviously be indexed by i, $\Pi_1, \Pi_2 \dots \Pi_N$.

When all agents randomly assign items to positions, Π is uniformly randomly distributed and exchangeable. This serves as our null hypothesis. The alternative is that a non-empty subset of agents manipulate their orderings in accordance with some *preference criterion*. In the targeted tests presented by GPG, this preference criterion is assumed to be uniform across agents and known by the analyst a priori; neither is assumed in the untargeted tests introduced here. For these tests, the alternative is simply that a subset of agents manipulate their orderings in accordance with their preferences, whatever they may be.[1] Any sufficient deviation from randomness, of any type, is taken to be evidence of manipulation of the orderings by a subset of agents. Section II will classify some different types of preference criteria and indicate which tests best suit each.

We discuss and ultimately compare all new or existing tests that examine all items contained in the ordering. These tests fall into three groups: the targeted tests developed by GPG, *aggregated* untargeted tests that rely only the cross-tabulation of items and their positions, and *disaggregated*, micro-level untargeted tests that utilize the individual orderings.

[1] Implicitly, these preferences factor in any order effects on decision-making. Thus, when there are primacy (recency) effects, the agent would tend to list their preferred items first (last). This point is inconsequential for the purposes of this section.

Targeted Tests. The two targeted tests developed by GPG are the Rank Compatibility Test and the Linear Concordance (LC) Test.

The Rank Compatibility Test. This test determines whether orderings that are *fully compatible* with posited preferences, ties being allowed, are unusually frequent. For example, let preferences over items A-F be A ~ B $\succ$ C ~ D ~ E $\succ$ F, where $\succ$ indicates preference and ~ indifference. Then the ordering {A, B, E, D, C, F} would be fully compatible, while the ordering {A, E, B, D, C, F} would not. Given any preference criterion, it is easy to determine the expected frequency of fully compatible orderings and then determine whether their observed frequency is unusually large.

The Linear Concordance Test. This test is sensitive not only to orderings that are fully compatible with posited preferences, but also to those that are *partly compatible*, such as {A, E, B, D, C, F} in the example above. As GPG note, such orderings would occur if the positions were filled sequentially and "each item's selection probability is an increasing function of perceived social status or a decreasing function of some discriminatory trait. [They] could also occur because of minor differences in biased agents' preferences...or other small frictions in the generation of a biased ordering from a preference criterion" (p. 6).

This test computes the mean concordance of a prescribed score vector, s ≡ (s_1, s_2, ... s_K), with the positions of each item in a set of orderings Π_i, i = 1..N. For *admissible* scores that have a mean of zero and a sum of squares of one, GPG show that this product, suitably scaled, asymptotically has a standard normal distribution under the null:

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} s_k \pi_{i,k} / \sqrt{NK(K+1)/12} \overset{a}{\sim} N(0,1) \qquad \textbf{(1)}$$

where L is the scaled linear concordance. When a qualitative preference ranking is posited, so that the strength of these preferences is unknown, GPG show that the optimal scores are the scaled, demeaned ranks (averaged over ties).

Aggregate Untargeted Tests. There are also two of these: the Max LC Test, which is new to the literature, and Ho and Imai's Rank Test.

The Max Linear Concordance Test. This test builds on the LC (linear concordance) test above, by simply searching for the *feasible* score vector that yields the largest value of L, called L^*.

There are two ways that "feasible" can be defined in this context. The first possibility is that preferences are merely qualitative; the agent can rank items but remains agnostic on the strength of those preferences (as in the A-F example above). For each possible preference ranking, the optimal score vector is determined as described above and the value of L determined. Selecting the largest of these across all possible preference rankings yields the L^* associated with a given set of orderings.

In this *strictly untargeted* approach, the set of feasible score vectors is finite, limited to the size of the set of possible preference rankings. This set grows rapidly in K.[2] Including ties, there are 75 possible preference rankings when K=4 and 4,683 when K=6. In this latter case, the set of feasible scores forms a fine grid spread evenly throughout the admissible portion of "score space."

In such cases it is just as well and far simpler computationally to ignore the discretization, allow all admissible scores to be feasible, and solve directly for the optimal score vector s^*. This approach, called *freely untargeted*, also applies when one is willing to quantify the strength of

[2] The set has $\sum_{k=1}^{K} k!S(K,k)$ elements, where S is the Stirling number of the second kind.

preferences. The program needed to solve for s* is simple–maximizing a linear objective with a quadratic constraint–and the solution equally simple (see the Appendix). The vector s* is determined, within a multiplicative constant, by the following formula:

$$s_k^* = \sum_{p=1}^{K} pC_{k,p} - N(K+1)/2 \tag{2}$$

where C$_{k,p}$ is the number of times item k has been placed in the pth position. One then normalizes the scores to have a 2-norm of one and calculates the associated value of L*. In either case it is straightforward to determine the distribution of L* under the null using Monte Carlo simulation.

The Rank Test. This test is based on the mean absolute difference in ranks between any two items. The test statistic R is calculated as follows:

$$R = \frac{\sum_{j \neq j^*} \frac{1}{N} \left| \sum_{i=1}^{N} (\pi_{ij} - \pi_{ij^*}) \right|}{K(K+1)/2} \tag{3}$$

where π_{ij} is the position, or rank, of item j in ordering i. The p-value is determined through Monte Carlo simulation.

This test is closely connected to the Max LC Test, because, per eq. **(2)**, the optimal score vector in the freely untargeted version of that test consists of the scaled mean ranks. To get the associated value of L*, this score vector is multiplied by the ranks themselves, which means that L* is a function of the squared mean ranks. We should expect similar p-values across the two tests.

Disaggregated Untargeted Tests. There are two of these, both new to the literature: the Equality of Permutations Test and the Cascading Chi-Squared Test.

Equality of Permutations Test. This test simply examines whether each permutation in the sample occurs with equal likelihood. When K = 2, there are only two possible permutations, so the binomial test of equal proportions applies. For K ≥ 3, we use the basic chi-squared test of equal proportions:

$$t^{E} = \frac{\sum_{p=1}^{K!} (C_p - N/K!)^2}{N/K!} \sim \chi^2(K!-1) \qquad \textbf{(4)}$$

where p = 1...K! indexes the set of all possible permutations and C_p is the count of observed occurrences of permutation p in the sample.

This test is simple, straightforward, and flexible. However, the number of permutations grows quickly in K, mandating a large sample size, N, in order to satisfy the standard requirement for a chi-squared test that all expected frequencies be at least 5, i.e., N/K! ≥ 5. Size calculations let us relax this rule somewhat: for K ≥ 4, a guideline of N/K! ≥ 2 works well in practice, perhaps because of the large number of permutations that are checked.[3] Nonetheless, even this looser guideline requires well over a thousand orderings when K is only 6.

The Cascading Chi-Squared Test. This test generalizes the well-known chi-squared test of equal proportions so that it can be applied to orderings (though in a different way than above).

Obviously, one can count the frequency that a given item, item k, is placed in 1st position, 2nd position, etc., in a set of orderings and calculate the appropriate chi-squared statistic, namely:

[3] When N = 2K!, simulated size at α = .01, .05, .10 is {.008, .034, .089} for K = 3, {.012, .045, .086} for K = 4, {.010, .050, .095} for K = 5, and {.009, .049, .099} for K = 6.

$$t^k = \frac{\sum_{p=1}^{K} (C_{k,p} - N/K)^2}{N/K} \sim \chi^2(K-1) \qquad \textbf{(5)}$$

where p = 1...K indexes positions and $C_{k,p}$ is the count of observed occurrences of item k in position p. One could do this for all items k = 1...K. However, the position of one item in an ordering affects the position of the others, so these placements are not independent. One could calculate t^k for each item, but they could not be meaningfully combined.

Our response is to adapt this approach so that the chi-squared statistics are independent. After calculating the chi-squared test statistic for one item in a set of orderings, simply "remove" that item from the set and analyze the order of the remaining items only. For example, in the A-F example above, count the frequency that A is placed in 1st position, 2nd position, etc., and calculate t^A, which is distributed $\chi^2(5)$ under the null. Then remove item A from all orderings, so that {A, E, B, D, C, F} becomes {E, B, D, C, F} and so on, and repeat the computation for item B. Under the null, that test statistic, t^B, is distributed $\chi^2(4)$ and is independent of its predecessor. Continuing this process to completion and summing the t^k's that result yields a single test statistic for the randomness of the overall ordering, which is asymptotically distributed $\chi^2(K(K-1)/2)$.[4]

II. Simulations.

[4] The final test statistic is not impervious to the order in which the items are selected for analysis. Thus, proceeding through items A-F in, say, reverse alphabetical order will not generally yield an identical result. For this reason, it is important to proceed through the items in a pre-ordained order, such as alphabetical order or numerical order, so as to prevent p-hacking. This approach is used in the applications below.

Setup. We wish to determine the power of each test as a function of 1) the number of items being ordered, K, 2) the number of orderings, N, and 3) the "strength" of preferences. We parameterize the latter so as to permit orderings that loosely correlate with preferences. That is, each ordering need not be fully compatible with the posited preference criterion, though the average position of each item in a large set of orderings will so conform.

Therefore, we generate the orderings using the following procedure. Each ordering is generated sequentially, selecting the next item from the remaining items with relative probabilities set as follows: $P_{k+m} / P_k = \delta^m$, $\delta \leq 1$, where k and m are positive integers, $k+m \leq K$, and P_j is the probability of selecting item j. The term δ represents the relative probability of item j being chosen relative to item j-1. When $\delta = 1$, there are no preferences and each ordering is randomly determined. When $\delta < 1$, item 1 $\succ$ item 2 $\succ$ item 3, and so on, with these preferences becoming stronger as δ falls. Targeted tests anticipate that this preference criterion governs any deviations from randomness, while untargeted tests do not.

Table A1 in the Appendix presents cross tabulations for 10,000 orderings when K = 4, using the four values of δ that are employed in our simulations: 0.95, 0.9, 0.8, and 0.7. It is apparent that $\delta = 0.95$ represents weak preferences, in which the first item appears in first position just a bit more than its confederates, while $\delta = 0.7$ represents preferences that are very strong. Thus our simulations encompass the full range of preference strengths.

For K and N, we use a range of values intended to represent the wide variety of sets of orderings to be found in the real world. When K = 2, the binomial test of equal proportions applies and the tests in this paper are unnecessary, so we choose $K \in \{3, 4, 5, 6, 10, 15, 20\}$ and $N \in \{50, 125, 250, 1000\}$. The values of K and N fall into these ranges for several of the applications presented below. (In a few cases, computational costs are O(K!) and quickly explode, so only the smaller values of K are used.)

Tables 1-4 present power at $\alpha = 0.05$, using 10,000 simulations, for all six tests considered in this paper, grouping the untargeted Max LC and Equality of Permutations Tests with their targeted equivalents, the LC and Rank Compatibility Tests. Overall, rejection rates range from 0.05 to 1, increasing as K and N grow and δ falls, as expected. Accordingly, when power equals one for a given combination of K, N, and δ, the entries for larger values of K and N and smaller values of δ are omitted, as they also implicitly equal one.

Results. We begin with the Rank Test, found in Table 1. This test exhibits very low power when preferences are mild and K and N are reasonably small. However, power grows steadily in both K and N. With few items, one can expect to reject the null hypothesis in small samples (N=50) when preferences are strong ($\delta \leq 0.8$) and in larger samples when preferences are weak. With many items ($K \geq 10$) rejection is likely even with small samples and weak preferences.

Next consider the Max LC test, presented in Table 2. The three panels of the table contain the three versions of this test: strictly untargeted, in which the only feasible score vectors are scaled, demeaned ranks; freely untargeted, in which all admissible score vectors are feasible; and targeted, that is, GPG's LC Test, which anticipates that item 1 is preferred over item 2, and so on.

We first compare the freely untargeted test to the Rank Test. Rejection rates are very similar across the two. This outcome is to be expected, since both tests are based on mean ranks, as mentioned above. Rejection rates are also similar for the strictly untargeted test, which is far more cumbersome to execute. The power of all of these tests, however, pales in comparison to that of the targeted LC test, presented in the last panel of Table 2. As in GPG, knowing the preference criterion that agents use in manipulating orderings is very helpful.

We now shift from aggregated tests to disaggregated tests. Results for the first of these, the Cascading Chi-Squared Test, are found in Table 3. There is a modest reduction in power relative to the two preceding tests, but the test can still detect deviations from randomness in small samples with strong preferences or large samples with weak preferences.

Finally, in Table 4, we consider the most flexible and straightforward test of all: the Equality of Permutations Test. Again, both untargeted and targeted versions are presented; the targeted version devolves to GPG's Rank Compatibility Test. Here, K does not exceed five, so as to satisfy the requirement that $N/K! \geq 2$.

The untargeted test modestly retreats from the power of the Cascading Chi-Squared Test, though it still retains the ability to detect deviations from randomness when samples are large or preferences are strong. Here, however, the targeted test generally performs worse, not better. There is an intuition behind this surprising finding. The targeted test puts all its eggs in one basket, so to speak, examining the frequency of just that single permutation that is fully compatible with the preference criterion. The untargeted test does not, and picks up the disproportionate number of orderings that loosely, but not perfectly, reflect preferences. This usually increases its power beyond that of the targeted test.

Discussion. The preference criterion we have adopted (when $\delta < 1$) is what GPG calls *unidirectional*. All agents who manipulate the orderings do so in accordance with the same hierarchical preferences, with only "random" or unsystematic deviations permitted. However, other possibilities exist.

This is most easily seen in the political context. There, unidirectional preferences would obtain if one candidate, say an incumbent, was generally more popular or highly regarded than his or her opponents, and orderings were influenced accordingly. An alternative would be *bidirectional* preferences that were based on ideology, not popularity. Any agent conducting an ordering could prefer either end of an ideological spectrum to the other, for example, may prefer liberals to conservatives or vice versa.[5] In this case, candidates with more extreme ideologies should be clustered at the highest and lowest positions in the orderings, with more moderate candidates in the middle. Widespread manipulation of the orderings may occur even though each item's mean position is similar, as liberals' and conservatives' manipulations cancel each other out.

Another, more complex alternative is that agents themselves are arranged on an ideological spectrum and prefer candidates with similar ideologies. In this case distinct, somewhat idiosyncratic sets of orderings will occur disproportionately, with some agents placing liberal candidates more highly, others placing conservatives highly, and others favoring moderates.[6] Manipulation of the orderings may occur even though the overall cross tabulation is quite balanced.

Such circumstances are not contemplated in our simulations, though they may occur in reality; it would be impractical to simulate every type of preference criterion in any event. Instead,

[5] Within the Republican party, ideology could manifest along a spectrum of Tea Party Republicans to "establishment" Republicans; within the Democratic party, it could manifest along a spectrum of Progressives to Centrists.

[6] To be more specific, orderings that "hopscotch" across ideological neighbors are less likely, because no agent should "rationally" create them, no matter what their ideology is. For example, let candidates be arranged as follows: $A \sim B \succ C \sim D \sim E \succ F$, where $\succ$ means "more conservative than" and ~ means ideological equivalence. No agent would intentionally construct the ordering {B, E, F, A, D, C}, so this ordering should be relatively infrequent.

we simply acknowledge that some of the four tests we consider "focus on" unidirectional preferences and some do not. Those that do will have more power to reject the null when preferences are, indeed, unidirectional, and less power when they are not. The former camp includes the Rank Test and Max LC Test. These tests rely on the mean position given to each item, which will deviate from each other the most under unidirectional preferences.

The Cascading Chi-Squared Test is a step removed from these two. It would be ideal for the bidirectional alternative contemplated above, as it should detect an uneven spread of items across positions, even when the mean rank is the same. The final test, the Equality of Permutations Test, is the most general of all, and will reject for any (sufficient) deviation from randomness, no matter what type. It is ideal for the second alternative contemplated above, in which manipulation may occur even though items are evenly spread across positions in the overall cross tabulation. However, this flexibility comes at a cost: this test absorbs many more degrees of freedom than its competitors.

In summary, there is a tradeoff between power and robustness. Tests that accommodate a wider variety of deviations from randomness will also have less power to reject the null under unidirectional preferences, and vice versa. The simulation results should be interpreted accordingly.

III. Applications.

The three applications to which we now apply these tests involve the ordering of candidates on primary election ballots, the order in which balls are drawn in Powerball, and the other in which contestants perform on *American Idol*. The first of these has practical relevance and permits a comparison of methods. The other two highlight the range of questions such tests can address and the implementation issues that can occur in doing so.

Primary Election Ballots. In primary elections in Texas, West Virginia, and at least one other state (Florida), candidate order is to be randomized at the county level by law, generating the repeated orderings needed to conduct our empirical tests. Applying our tests to this data will indicate whether this law was followed uniformly across all counties in each of these states or whether the orderings were manipulated in some counties in order to advantage or disadvantage certain candidates. The ballot order effect implies that candidates gain from being listed higher on the ballot, and in our experience this fact is widely understood at the local level.

For the last two electoral cycles, 2018 and 2020, county-level ballot orders are available online from Texas' and West Virginia's Secretary of State. A total of 41 statewide contests were held in these states in those years, allowing us to amass a large body of evidence and facilitating comparisons across parties, states, and type of contest (executive, legislative, judicial).

Of these 41 contests, 17 have just two candidates and utilize the binomial test of equal proportions, while 24 have at least three candidates, necessitating the tests analyzed here. For these contests, we utilize all four such tests, except in a few races in which the Equality of Permutation Test's requirement that $N/K! \geq 2$ is not satisfied. We can thus compare the performance of these four tests in real-life scenarios that may or may not satisfy the unidirectional preferences posited in our simulations.

The strength of preferences and the values of K and N also vary across these contests. Texas has 254 counties, West Virginia just 55. Many multi-candidate contests have just three or four competitors, but some others, such as those for president, have more than ten. Finally, the desire to manipulate the orderings is likely to be stronger in high-profile contests such as those for president and governor than it is in, say, judicial races–especially in West Virginia, where judicial contests are non-partisan (yet, fortuitously, decided during these elections using the same randomization scheme).

The results are presented in Table 5, broken down by state, year, and type of contest. Each non-empty cell has four rows, each containing p-values. The top row pertains to the Equality of Permutations Test–which devolves to the binomial test of equal proportions when K=2, and is so indicated by italics. When K ≥ 3, the second row presents p-values from the Cascading Chi-Squared Test, which can be compared to the respective non-italicized values in the row above it. The third and fourth rows pertain to the remaining tests, the Rank Test and the Max LC test, which again can be compared to those above them within the same cell. Thus, every vertical column of unitalicized numbers within a cell refers to multiple estimates for a single race (for the West Virginia Supreme Court in 2020, the three vertical columns contain four estimates for each of three races). If the cell is empty, there was no election for that office in that state in that year or the (primary) contest had just one candidate.

Turning first to Texas, on the left side of the table, two contests have a straight run of .00 p-values: the 2018 Democratic primary for U.S. Senate, in which popular candidate Beto O'Rourke was disproportionately listed first, and the 2020 Democratic primary for Railroad Commissioner, in which eventual third-place finisher Kelly Stone was disproportionately listed higher on the ballot. In these two contests, there is clear evidence that some orderings were manipulated so as to place these candidates higher on the ballot. In five other races, both Democratic and Republican, the p-values are "suggestive," hovering in the vicinity of .10. In the remaining contests, the p-values–including those from the binomial test–are widely distributed across the unit interval, as would be expected if candidate order is randomly determined in all counties.

Across the 28 Texas contests analyzed, random chance should generate three for which $p \leq .1$ and five or six for which $p \leq .2$. Depending on which test is considered, the actual number of contests meeting those criteria exceeds the expected number by about four. Overall, then, ballot order manipulation is occasionally present in recent Texas primary elections, but is not widespread; when it occurs, sometimes several counties are involved (when p = .00, see below) and sometimes just a few (generating the small cluster of p-values near .10). These findings are more sanguine than Grant's (2017) for the 2014 Texas primaries, in which manipulation occurred at a higher rate. The difference may be explained by the enactment of a new state law requiring county chairs to report ballot order to Texas' Secretary of State, who then publishes them on the internet. This sunlight may serve as a disinfectant.

Turning to West Virginia, the findings are quite different. The evidence for ballot order manipulation is both weaker and more prevalent. Small p-values such as .00 or .01 are not to be found, because N is only 55, but seven of eleven partisan races have p-values below .20, far greater than the two that would be expected from random chance. This finding indicates that ballot order manipulation is reasonably common in West Virginia, though it is hard to pinpoint the amount that occurs in any given race. The main exception to this rule occurs in the judicial elections, where the p-values are relatively high. This may be a consequence of the non-partisan nature of these contests in that state. In Texas, where judicial elections are partisan, low p-values are not infrequent.

We now consider the strength of the different tests for randomness. The tests in the top two rows of the table, the Full Permutation and Cascading Chi-Squared Tests, are more flexible but less powerful when preferences are unidirectional. It is the other way around for the tests in the bottom two rows, the Rank and Max LC Tests. How do their results compare?

Broadly speaking, the p-values tend to be reasonably similar. In Texas, especially, no one test clearly and consistently stands out above the others. When there is strong evidence of manipulation, the p-values are generally low across the board. Things are more variable when the p-values are only suggestive, but only slightly so. This suggests that preferences are reasonably unidirectional, though not so much that the Rank and Max LC Tests are clearly superior. The fact that the candidates benefitting from ballot order manipulation tend to be popular incumbents or high-profile challengers supports this general finding of unidirectional preferences.

The most striking exception to this rule occurs not in Texas, but in the 2020 Democratic primary for U.S. Senate in West Virginia. There, the two more flexible tests each yield p-values of about .05, while the two less flexible tests yield p-values of about .55. It seems counterintuitive that two groups of tests should reach such divergent conclusions, but they do, and the reason why is instructive. This race featured three candidates, two of which (Robb and Swearengin) were each listed first and last a disproportionate number of times, with the remaining candidate (Ojeda) disproportionately listed in the middle position. As a result, each candidate's mean rank was almost exactly 2.0. The Rank and Max LC Tests, powerless to detect these offsetting, bidirectional preferences, returned large p-values; the other, more flexible tests easily detected this manipulation.

A closer look at West Virginia's results yields several similar, though milder, instances of this phenomenon: in the 2020 Republican contests for president and U.S. Senator and in one Supreme Court race. Unlike Texas, it does appear that bidirectional preferences are genuine in West Virginia, favoring the more flexible Cascading Chi-Squared and Equality of Permutations Tests. The reason why is also instructive. In Texas, the county chair of each party has the sole responsibility for conducting the orderings, so Republican candidates are ordered by Republican county chairs, and the same for Democrats. In West Virginia, all orderings are conducted by the county clerk, who is elected in a partisan contest. This means that the same set of candidates is ordered in some counties by a Republican and in other counties by a Democrat. It is not surprising that their preferences are reversed.

Finally, these data contain one instructive instance of targeted vs. untargeted testing: Texas' 2020 Republican presidential primary. Overall, the untargeted p-values for this race, .02, .09, and .14, merely suggest the presence of manipulation, but with seven candidates in this race, these tests may lack power. However, prior evidence for Texas 2014 primary and runoff elections (Grant, 2017) indicates that such manipulation tends to favor well-known candidates, often incumbents, who are popular within the party. This implies any such manipulation would favor the incumbent, Donald Trump. This supports a targeted test that is potentially much more powerful.

Accordingly, we applied GPG's targeted LC and Rank Compatibility Tests to this data, using the alternative hypothesis that Donald Trump was preferred to all other candidates (between whom no preferences existed). The results were indeed far stronger, rejecting the null in favor of this alternative with p-values of .00 for both tests. Substantial manipulation accompanies such low p-values. Trump was listed first on 52 counties' ballots, when the expected incidence was 254/7 = 36.3. In the Democratic Senate primary mentioned earlier, Beto O'Rourke was listed first on 118 counties' ballots, when the expected incidence was 254/3 = 84.7. Roughly one-third of counties listing Trump first on the ballot did so deliberately, as did roughly one-quarter of counties listing O'Rourke first. Given the size of the ballot order effect in primary elections, this degree of manipulation could easily sway a close contest. Fortunately, these elections weren't close.

Powerball. Powerball, the most famous lottery in the U.S., is conducted semi-weekly by the Multi-State Lottery Association, with tickets sold in 45 states and some U.S. territories. Held since 1992, the lottery draws five numbered "white balls" plus a red numbered "Powerball." To win the jackpot, which can amount to hundreds of millions of dollars, the ticket must match the numbers on all six balls. These balls are drawn physically and sequentially using mechanical devices that resemble popcorn poppers; years' worth of drawings can be viewed online. The Wisconsin Lottery's web site had posted over fifteen years' of winning numbers in the order that they were drawn, a total of 1,596 drawings. We use this data for our analysis.

To test the null hypothesis that these balls are drawn randomly, one could apply the chi-squared test of equal proportions to the number of times each ball is selected in these drawings. However, the only way a ball would be drawn "too much" would be for it to have a mechanical flaw that caused it to edge out other balls for selection. If so, it should also be drawn earlier in the sequence, on average.[7] Thus, with sufficient data, a test of the randomness of the *orderings* should outweigh one that merely examines *frequencies*, since it incorporates information on the sequence in which the balls are drawn. The appropriate test is untargeted, as there is no reason to expect any deviations from randomness to take a particular form.

This problem, though intriguing, is messy. One complication is minor: the number of balls used in the drawing has not been constant over time. Over our sample period, Feb. 18, 2004 - June 1, 2019, the number of white balls increased from 53 to 55 to 59 to 69; the number of red balls,

[7] This logic is analogous to the simulations in the previous section, where smaller numbers had a higher probability of being selected for the next position in the sequence. Therefore, any deviations from randomness are expected to be unidirectional. However, the circumstances do not permit use of the tests best suited to these kinds of "preferences," as will be demonstrated shortly.

drawn using a separate machine and not analyzed here, also increased. The other complication is major: only a subset of white balls are ordered each week–those five that are selected by the "popcorn popper." The remainder are unordered. We call this data structure *partly blocked*: the same number of items are ordered each week, but they are only a subset of all items. All of the methods in this paper are designed for *fully blocked* data in which every item is ordered. How can we test the randomness of incomplete orderings such as these?

The solution is to adapt the Cascading Chi-Squared Test so that it runs "vertically" rather than "horizontally." Rather than proceeding by *item*, examining at the positions of the 1 ball across all orderings and so on, we proceed by *position*. Initially, we count the number of 1's, 2's, etc., that are drawn first in the set of orderings and calculate the appropriate chi-squared statistic. Then we count the number of 1's, 2's, etc., that are drawn second and calculate a second chi-squared statistic in the usual way, adjusting the expected number of 1's, 2's, etc. to account for those that have already been drawn (for first position). We then continue this process for all five positions that are drawn and sum these five statistics.

This approach bastardizes the usual chi-squared test: for every position after the first, the assumption of independence is not strictly satisfied. It turns out, however, that when only five numbers are drawn from a much larger set, any deviation from the traditional distribution is small.[8] Nonetheless, to be thorough, p-values are calculated using Monte Carlo simulation.

The results are placed in the first column of Table 6. As K increased three times over the

[8] For fully blocked data, these statistics' sum has a distribution that closely resembles the chi-squared, with an expected value of $K(K - H_K)$, where H_K is the Kth harmonic number; the test is slightly undersized. This problem is smaller for partly blocked data, as the non-independence problem is "weaker." This test has less power than the Cascading Chi-Squared Test for fully blocked data, so it is useful only with partly blocked data, as here.

sample period, estimates are presented for three subperiods, in which K is 55, then 59, then 69. Following this is the estimated p-value for the full sample period. Our estimation approach easily handles the variation in K over this period, by simply adjusting the expected number of occurrences of each ball accordingly.

The p-values are generally large; there is no evidence that the drawings are non-random. The results are similar for the counts, analyzed with the chi-squared test of equality of proportions in the second column of the table. The p-values are generally similar across the two tests, with no clear sign of superiority of either. This is as expected, as one would not expect manipulation to occur in a highly visible, mechanized process in which the incentives run in the other direction.

American Idol. The popular television show *American Idol* ran for fifteen seasons on the Fox television network, from 2002-2016.[9] Each season pares down an initial set of roughly twelve finalists to a single winner, generally by eliminating one finalist each week. Page and Page (2010) have shown that contestants' success is influenced by the sequence in which they perform on the program, with earlier-performing singers receiving fewer votes from the television audience and exiting at higher rates.

There is little evidence on how this sequence of contestants is determined each week. It appears not to be literally randomly determined, though the process may still be as good as random for practical purposes.

> Producers decide the singing order except for the finale, which is a singer's choice after a coin flip. They vary the order each week to be fair, but also try to arrange

[9] The show has since moved to a different network. As this other network could use a different procedure for ordering the contestants, its seasons of the show are not analyzed.

> singers and their songs to make the most entertaining show, executive producer Ken Warwick says. "It's worked out with two things in mind: where the kid (performed) last week, and if it's a slow, 'dirgey' ballad, I try not to open with that," he says. ("'Idol' Singers Who Go First May Not Last," *Norwich Bulletin*, Apr. 21, 2008).

Broadly speaking, then, there are three possibilities:

- The orderings are as good as randomly determined.
- Certain performers are systematically placed earlier (later) in the program, perhaps because they regularly perform upbeat (downbeat) songs.
- The orderings are conducted so as to even out imbalances over time, that is, to "correct" random inequities in earlier weeks, out of concern for fairness.

Because we have data on multiple seasons of the show, each of which has many weeks of orderings, we can statistically determine which of these statements best reflects reality. Our statistical test yields a p-value for each season. If these p-values are widely spread across the unit interval, the first possibility is most credible. If they cluster below .5, the second possibility is most credible. If they cluster above .5, so that there is less variation than is implied by random chance, the third possibility is most credible.

Conducting these tests poses a problem, however: the data are not in a block design at all. Typically, one candidate is eliminated each week. Thus, if a season's first week contains twelve contestants, the second week contains eleven, and so on. None of the tests in this paper directly apply to this *unblocked* data structure. In order to analyze the randomness of these orderings, it is again necessary to adapt one of these tests to suit this purpose.

The best candidate is the Max LC Test. As discussed above, this test has two variants: freely untargeted and strictly untargeted. In the latter, the score vector can only take a limited set of values: those that are optimal under some preference ranking, assuming that the strength of these preferences is not known. For any such hypothesized preference ranking, the optimum score vector is the hypothesized ranks, scaled to have a mean of zero and a 2-norm of one. The feasible set contains all such admissible score vectors.

To adapt this test to this data, we assume that the elimination of any one contestant does not change the preference ranking over those that remain. If Contestant 1 $\succ$ Contestant 2 $\succ$ Contestant 3 $\succ$ Contestant 4, and Contestant 3 is eliminated that week, then the next week's preferences are Contestant 1 $\succ$ Contestant 2 $\succ$ Contestant 4. Then, given any initial preference ordering, one can calculate the concordance L in equation **(1)** for any subset of items, simply by modifying K and the score vector accordingly. Doing so for each week of the season, one can calculate the mean concordance across the season for any and all initial preference orderings. The modified Max LC statistic is associated with the hypothesized preferences that yield the largest mean concordance. The p-value associated with this statistic can be calculated through Monte Carlo simulation.

This is simple enough in theory, but in practice the number of possible preference rankings grows quickly in K. Allowing indifference between contestants, that is, ties in these rankings, there are over 1.6 billion possible preference rankings when K = 11. As this application is merely illustrative, we simplify matters by limiting K to 10 and excluding preference criteria that allow indifference between contestants. Accordingly, we analyze the last nine weeks of the ten seasons of *American Idol*, listed in Table 7, that eliminate exactly one contestant per week over that period (or close enough to it that any deviations can be finessed, as explained in the note to the table).

The p-values for these ten seasons, reported in this table, are widely disbursed across the unit interval, with a mean of .46. This clearly supports the first possibility articulated above. In the Fox years of *American Idol*, the contestant orderings were as good as randomly conducted. In their analysis of order effects on success probabilities in *American Idol*, Page and Page (2010) did not detect any bias in their estimates arising from non-random sequencing. Their finding is consistent with ours.

IV. Discussion and Conclusion.

The three applications just presented–ballot order, Powerball drawings, and the sequence of contestants on *American Idol*–displayed tremendous variation in several respects.

- **Parameter Values.** The number of items being ordered, K, ranged from 2 to 69, while the number of orderings, N, ranged from 9 to 1,596. The statistical power of tests for randomness varied accordingly.

- **Data Structure.** Ballot orderings were fully blocked, so that each observation was an ordering of all items; Powerball was partially blocked, so that observation was an ordering of some, but not all, items; *American Idol* was unblocked, so that the number of items being ordered changed from observation to observation.

- **Computational Issues.** While most tests were so simple that they could be (and sometimes were) executed in Excel, this was not so for *American Idol*, which was computationally intensive.

- **Strength of Preferences.** Preferences were sometimes strong, as in a few primary contests in Texas, sometimes weak, as in other primary contests in that state and in West Virginia, and sometimes non-existent, so that the null hypothesis of full randomization was not rejected at even suggestive α levels.

- **Type of Preferences.** In Texas primary elections, especially, the underlying preference criterion was unidirectional, but in some primaries in West Virginia, it was bidirectional. Institutional factors explained the difference.

- **Underlying Mechanism.** Orderings might deviate from random because of mechanical flaws (Powerball), self-interest (ballot order), or concern for others (*American Idol*).

- **A Priori Information.** It was generally impractical to anticipate the preference criterion that would govern any deviations from randomness; however, there were some exceptions.

This variation means that there is no single best test for the randomness of repeated orderings. When there is an adequate basis to specify a preference criterion a priori, targeted tests are preferred; otherwise, untargeted tests must be used. When preferences are unidirectional and power is limited, the Rank Test and Max LC Test are preferred. When preferences may not be unidirectional, so robustness is a concern, and N is large relative to K, the Cascading Chi-Squared and Equality of Permutation Tests prevail. When the data are not fully blocked, it may be necessary to adapt the Cascading Chi-Squared or Max LC tests to suit, sometimes increasing computational costs in consequence. And when an agent conducting repeated orderings cares about fairness, he or she may deliberately offset earlier, random deviations in sequencing, generating high p-values, not low ones. All of these considerations appeared in our applications.

To suit these varied circumstances, this paper has considered six tests for the randomness of orderings–three old, three new–and has demonstrated how these tests can be adapted to suit a particular data set when necessary. We hope this toolkit will equip researchers to examine the randomness of repeated orderings in the wide variety of places in which they occur in everyday life.

REFERENCES

Antipov, E., and E. Pokryshevskaya. 2017. Order effects in the results of song contests: Evidence from the Eurovision and the New Wave. *Judgment and Decision Making* 12, 415-519.

Bruine de Bruin, W. 2006. Save the last dance II: unwanted serial position effects in figure skating judgments. *Acta Psychologica* 123, 299-311.

Carney, D., and M. Banaji. First is best. *Plos One* 7:e35088 (2012).

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, 166-178.

Dean, M. 1980. Presentation order effects in product taste tests. *Journal of Psychology* 105:107-110.

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* 25,109-129.

Grant, D. 2017. The ballot order effect is huge: evidence from Texas. *Public Choice* 172, 421–442.

Grant, S., M. Perlman, and D. Grant. 2020. Targeted testing for bias in order assignment, with an application to Texas election ballots. *Journal of Statistical Planning and Inference* 206:12-28.

Haan, M., G. Dijkstra, and P. Dijkstra. 2005. Expert judgment versus public opinion–evidence from the Eurovision song contest. *Journal of Cultural Economics* 29:59-78.

Ho, D., and K. Imai. 2008. Estimating causal effects of ballot order from a randomized natural experiment: the California alphabet lottery. *Public Opinion Quarterly* 72:216-240.

Krosnick, J. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 213-236.

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

Miller, J., and J. Krosnick. 1998. The impact of candidate name order on election outcomes. *Public Opinion Quarterly* 62, 291-330.

Meredith, M., and V. Salant. 2013. On the causes and consequences of ballot order effects. *Political Behavior* 35, 175-197.

Mussweiler, T. 2003. Comparison processes in social judgment: mechanisms and consequences. *Psychological Review* 110, 472-489.

Novemsky, N., and R. Dhar. 2005. Goal fulfillment and goal targets in sequential choice. *Journal of Consumer Research* 32, 396-404.

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

Salant, Y. 2011. Procedural analysis of choice rules, with applications to bounded rationality. *American Economic Review* 101, 724-748.

Table 1. Power Simulations, Rank Test (power at α = .05 for δ = .95, .9, .8, and .7, in that order).

K↓ N→	50	125	250	1000
3	.07, .12, .33, .69	.07, .17, .63, .97	.10, .33, .93, 1	.31, .89, 1
4	.07, .16, .60, .96	.10, .35, .96, 1	.18, .66, 1	.64, 1
5	.09, .28, .89, 1	.16, .62, 1	.34, .94, 1	.92, 1
6	.14, .46, .99, 1	.26, .88, 1	.54, 1	.99, 1
10	.42, .99, 1	.87, 1	1	1
15	.93, 1	1		
20	1			

Table 2. Power Simulations, Freely Untargeted, Strictly Untargeted, and Targeted Max LC Tests (power at $\alpha = .05$ for $\delta = .95, .9, .8$, and .7, in that order).

Freely Untargeted

K↓ N→	50	125	250	1000
3	.06, .09, .29, .63	.08, .18, .67, .96	.11, .32, .93, 1	.31, .90, 1
4	.07, .16, .61, .96	.11, .36, .96, 1	.18, .67, 1	.65, 1
5	.09, .27, .89, 1	.17, .63, 1	.33, .93, 1	.93, 1
6	.12, .44, .99, 1	.26, .87, 1	.54, 1	1
10	.41, .99, 1	.87, 1	1	
15	.93, 1	1		
20	1			

Strictly Untargeted

K↓ N→	50	125	250	1000
3	.06, .11, .30, .64	.08, .19, .66, .98	.12, .35, .93, 1	.32, .90, 1
4	.08, .17, .64, .97	.11, .37, .97, 1	.20, .68, 1	.67, 1
5	.10, .30, .91, 1	.18, .68, 1	.33, .95, 1	.93, 1
6	.13, .46, .99, 1	.29, .90, 1	.56, 1	

Targeted

K↓ N→	50	125	250	1000
3	.11, .20, .50, .83	.14, .33, .83, .99	.22, .55, .98, 1	.54, .97, 1
4	.15, .35, .86, 1	.25, .64, 1	.40, .89, 1	.87, 1
5	.23, .58, .99, 1	.42, .90, 1	.65, .99, 1	.99, 1
6	.33, .80, 1	.60, .99, 1	.86, 1	1
10	.86, 1	1		
15	1			
20				

Note: the strictly untargeted test was very slow to run for $N \geq 10$ and so was omitted. Power for omitted entries is implicitly one, as simulations determined it to be one for less favorable values of N, K, and/or δ.

Table 3. Power Simulations, Cascading Chi-Squared Test (power at $\alpha = .05$ for $\delta = .95, .9, .8$, and .7, in that order).

K↓ N→	50	125	250	1000
3	.06, .09, .25, .57	.07, .16, .57, .95	.09, .28, .89, 1	.27, .85, 1
4	.07, .13, .47, .91	.09, .28, .91, 1	.14, .55, 1	.53, 1
5	.07, .18, .75, 1	.12, .46, 1	.22, .83, 1	.81, 1
6	.08, .27, .93, 1	.16, .69, 1	.35, .98, 1	.97, 1
10	.20, .82, 1	.55, 1	.93, 1	1
15	.49, 1	.98, 1	1	
20	.87, 1	1		

Table 4. Simulation Results, Untargeted and Targeted Equality of Permutation Tests (power at $\alpha = .05$ for $\delta = .95$, .9, .8, and .7, in that order).

Untargeted

K↓ N→	50	125	250	1000
3	.05, .07, .21, .51	.07, .13, .51, .93	.08, .24, .85, 1	.22, .80, 1, 1
4	*.06, .09, .28, .71*	.07, .16, .71, 1	.10, .32, .98, 1	.31, .97, 1, 1
5	----	----	*.10, .33, .99, 1*	.30, .99, 1, 1

Targeted

K↓ N→	50	125	250	1000
3	.05, .09, .23, .49	.08, .17, .49, .85	.11, .26, .74, .98	.29, .73, 1, 1
4	*.04, .06, .18, .43*	.08, .17, .50, .87	.10, .24, .75, .98	.27, .70, 1, 1
5	----	----	*.05, .14, .50, .92*	.19, .52, .98, 1

Note: Results in italics satisfy $2K! \leq N < 5K!$.

Table 5. P-values from Randomization Tests, Texas and West Virginia Primaries (in vertical order within each cell, p-values from the Equality of Permutation Test, the Cascading Chi-Squared Test, the Rank Test, and the freely untargeted Max LC Test).

	TEXAS				WEST VIRGINIA			
	2018		2020		2018		2020	
Office	Democratic	Republican	Democratic	Republican	Democratic	Republican	Democratic	Republican
President	----	----	-- .81 .33 .38	-- .14 .09 .02	----	----	-- .15 .24 .20	-- .15 .54 .56
U.S. Senate	.00 .00 .00 .00	.06 .34 .08 .10	-- .67 .54 .54	.48 .55 .44 .47	*1.00* --	-- .14 .09 .09	.06 .04 .54 .55	.17 .15 .31 .33
Governor	-- .70 .87 .88	.48 .66 .88 .92	----	----	----	----	-- .17 .14 .17	-- .03 .04 .03
Agriculture Comm.	----	.35 .30 .09 .11	----	----	----	----	.31 .25 .22 .21	*1.00* --
Other State Offices	*.85,.15,.23,.75* --	*.75*, *.12*, .04 .13 .29 .29	.00 .00 .00 .00	*.75* --	----	----	*.59* --	----
Supreme Court	----	----	*.95,.75,.49,.49*	----	----		.74, .81, .07 .58, .92, .32 .43, .41, .73 .47, .41, .74	
Court of Criminal Appeals	----	*.09*, .05 .12 .01 .02	*.57*, .11 .10 .08 .09	*.19* --	----		----	

Note: Each line contains p-values for all contests fitting the description for that cell. In West Virginia, Supreme Court positions are non-partisan races placed in the same ballot order on both parties' primary ballots. Supreme Court positions contested in Texas in 2018 were were Places 2, 4, and 6; in 2020, the Chief Justice and Places 6, 7, and 8; Court of Criminal Appeals positions contested in 2018 were the Presiding Judge and Place 8; in 2020 were Places 3 and 4. "Other State Offices" includes Attorney General in West Virginia and, in Texas, Lt. Governor, Comptroller, and Land Commissioner in 2018 and Railroad Commissioner in both 2018 and 2020. Italicized numbers involve binomial tests in two-candidate races. When the Full Permutation test is not presented, but other multi-candidate tests are presented, it is because the requirement N/K! ≥ 2 was not met.

Table 6. Randomization Tests of Powerball Drawings (χ^2 test statistic, with p-values in parentheses).

Time Frame	Modified Cascading χ^2 Test	Equality of Proportions Test
Aug. 31, 2005 - Jan. 03, 2009 (K = 55, N = 350)	245.1 (0.85)	46.2 (0.76)
Jan. 07, 2009 - Oct. 3, 2015 (K = 59, N = 704)	258.2 (0.90)	37.7 (0.98)
Oct. 7, 2015 - June 1, 2019 (K = 69, N = 382)	364.9 (0.16)	75.7 (0.24)
Feb. 18, 2004 - June 1, 2019 (K = 69*, N = 1,596)	307.9 (0.93)	63.4 (0.63)

* K = 53 from Feb. 18, 2004 - Aug. 27, 2005, and then grew as listed in the table.

Table 7. Results, American Idol, Modified Max LC Test.

	p-value(s)
Season 2	0.37
Season 3	0.15
Season 4	0.81
Season 5	0.38
Season 7	0.93
Season 10	0.47 / 0.49
Season 11	0.04
Season 12	0.51
Season 13	0.61
Season 14	0.30 / 0.37

Note: Season 2 removed Corey Clark, who was disqualified in an early round, from all rankings. In Seasons 10 and 14, there were 11 people in the first round analyzed, two of which were eliminated before the consequent round. For these seasons, the test was run twice, first retaining one, then the other, of these two contestants. In several seasons, there were earlier rounds with additional contestants; only the last nine rounds were analyzed.

Appendix

Define α_j as item j's sum of ranks, that is, the number of times it is ranked first, plus two times the number of times it is ranked second, and so on; i.e., $\Sigma(pC_{j,p})$, where p indexes position or rank. Then, the program to be solved can be written as follows:

$$Lagrangian = \sum_{i=1}^{K} \alpha_i s_i - \lambda_1 \sum_{i=1}^{K} s_i - \lambda_2 (\sum_{i=1}^{K} s_i^2 - 1) \qquad \textbf{(6)}$$

where λ_1 and λ_2 are Lagrange Multipliers associated with the constraints that the score vector s have a mean of zero and a 2-norm of one.

Taking the derivative with respect to s_j yields the following first order condition:

$$s_j^* = (\alpha_j - \lambda_1)/2\lambda_2 \quad \forall j \qquad \textbf{(7)}$$

Summing the s_j^* and setting equal to zero yields:

$$\sum_{j=1}^{K} s_j^* = \sum_{j=1}^{K} (\alpha_j - \lambda_1)/2\lambda_2 = 0 \quad \Rightarrow \quad \lambda_1 = \frac{1}{K} \sum_{j=1}^{K} \alpha_j = \bar{\alpha} \qquad \textbf{(8)}$$

One can easily show that $\bar{\alpha} = N(K+1)/2$. Thus, for all j, s_j^* is proportional to α_j - N(K+1)/2, as claimed in the text.

Table A1. Relative Frequencies of Positions by Item, K=4, for the δ Values Used in Tables 1-4.

δ = 0.95	First Position	Second Position	Third Position	Fourth Position
First Item	0.276	0.257	0.238	0.229
Second Item	0.260	0.253	0.249	0.237
Third Item	0.245	0.251	0.251	0.253
Fourth Item	0.219	0.239	0.262	0.280

δ = 0.90	First Position	Second Position	Third Position	Fourth Position
First Item	0.295	0.258	0.229	0.218
Second Item	0.268	0.255	0.255	0.222
Third Item	0.244	0.262	0.251	0.244
Fourth Item	0.193	0.225	0.265	0.317

δ = 0.80	First Position	Second Position	Third Position	Fourth Position
First Item	0.344	0.267	0.213	0.176
Second Item	0.288	0.271	0.238	0.204
Third Item	0.228	0.257	0.268	0.247
Fourth Item	0.140	0.208	0.282	0.374

δ = 0.70	First Position	Second Position	Third Position	Fourth Position
First Item	0.394	0.279	0.196	0.131
Second Item	0.311	0.282	0.234	0.173
Third Item	0.202	0.270	0.282	0.246
Fourth Item	0.093	0.169	0.289	0.449

Note: These relative frequencies are calculated from 10,000 simulations for each value of δ.