Mathematics, Statistics and Data Science
Faculty Works

Mathematics, Statistics and Data Science

2012

Linear rank tests of uniformity: Understanding inconsistent outcomes and the construction of new tests

Anna E. Bargagliotti
Loyola Marymount University, Anna.Bargagliotti@lmu.edu

Follow this and additional works at: https://digitalcommons.lmu.edu/math_fac

Part of the Education Commons, and the Statistics and Probability Commons

Taylor & Francis
Taylor & Francis Group

Linear rank tests of uniformity: understanding inconsistent outcomes and the construction of new tests

Anna E. Bargagliotti[a]* and Michael E. Orrison[b]

[a]*Department of Mathematics, Loyola Marymount University, Los Angeles, CA 90045, USA;* [b]*Department of Mathematics, Harvey Mudd College, Claremont, CA 91711, USA*

Several nonparametric tests exist to test for differences among alternatives when using ranked data. Testing for differences among alternatives amounts to testing for uniformity over the set of possible permutations of the alternatives. Well-known tests of uniformity, such as the Friedman test or the Anderson test, are based on the impact of the usual limiting theorems (e.g. central limit theorem) and the results of different summary statistics (e.g. mean ranks, marginals, and pairwise ranks). Inconsistencies can occur among statistical tests' outcomes – different statistical tests can yield different outcomes when applied to the same ranked data. In this paper, we describe a conceptual framework that naturally decomposes the underlying ranked data space. Using the framework, we explain why test results can differ and how their differences are related. In practice, one may choose a test based on the power or the structure of the ranked data. We discuss the implications of these choices and illustrate that for data meeting certain conditions, no existing test is effective in detecting nonuniformity. Finally, using a real data example, we illustrate how to construct new linear rank tests of uniformity.

Keywords: tests of uniformity; nonparametric; rankings; Friedman test; Anderson test; effective space

1.   Introduction

Spurred by Friedman's (1937) paper, several nonparametric tests have been proposed to address the problem of n rankings – the problem of determining whether m alternatives that have been fully ranked by a sample of n judges are significantly different. In each case, the null hypothesis of uniformity over the set of all permutations of the m alternatives is tested.

Widely used statistics to test for uniformity include Kendall's (1938) and Kendall and Smith's (1939) concordance statistic, Spearman's (1904) correlation coefficient, Cayley's (1849) distance statistic, Friedman's (1937) randomised block statistic, Anderson's (1959) marginals statistic, and the pairwise statistic associated with Wilcoxon's (1945) test. Several of these statistics (e.g. Friedman 1937; Wilcoxon 1945; Anderson 1959) can be realised as linear rank statistics as defined in Hajek, Sidak, and Sen (1999).

Interestingly, as noted in Marden (1995), different tests of uniformity can lead to conflicting results when applied to the same data set. For example, consider the rankings of 60 judges comparing three alternatives $\{A, B, C\}$ resulting in the following ranked data set:

Ranking	Number of judges	
ABC	6	
ACB	10	
BAC	6	(1)
BCA	10	
CAB	14	
CBA	14	

These data led the Friedman test (Friedman 1937) to reject the null hypothesis of uniformity, whereas using the Anderson test (Anderson 1959) leads to a failure to reject.

Building off the ideas found in Marden (1995, chap. 3), this paper explains why different linear rank tests of uniformity can yield inconsistent test outcomes. Our approach describes a conceptual and unifying framework for understanding well-known linear rank tests of uniformity while also allowing us to highlight practical guidelines and straightforward motivation for the construction of new linear rank tests of uniformity. As such, we believe our results will be of interest to theoreticians and application-driven researchers alike.

Section 2 introduces the framework to discuss the tests considered throughout this paper. Section 3 defines the class of linear rank statistical tests which Section 4 connects to a set of natural summary statistics computed from the ranked data. A natural decomposition of the data space characterising all inconsistencies among test outcomes is then presented in Section 5. Section 6 connects the decomposition to the power of a test. Sections 7 and 8 provide motivation for constructing new tests while Section 9 uses a real data example to show how to construct a new test. The last sections provide insights as to when certain tests are more desirable to use than others and discusses the implications and extensions of the choice of test.

2. Distributions defined on rankings

In this and in Section 3, we follow closely the ideas and notation presented in Marden (1995, chap. 3).

Suppose m alternatives $A_1, \ldots, A_m$ are fully ranked by n judges from the most preferred to the least preferred. Thus, if we denote the set of all possible rankings by $\mathcal{S}_m$, then each judge is being asked to choose a single element Y from $\mathcal{S}_m$. Furthermore, suppose these data are generated as n independent and identically distributed replicates of $Y \sim P$, where P is a probability distribution on $\mathcal{S}_m$.

We can encode the probability distribution P as a vector in $\mathbb{R}^{m!}$, where if y_i is the ith permutation in $\mathcal{S}_m$, then the ith entry of P is

$$P[Y = y_i].$$

The resulting data vector or *profile* $\mathbf{p}$ can also be encoded as a vector in $\mathbb{R}^{m!}$ where the ith entry of $\mathbf{p}$ is the number of judges who chose the ranking y_i.

For example, if $m = 3$, and the rankings of the alternatives are ordered lexicographically, then the profile

$$\mathbf{p} = [8, 16, 6, 18, 10, 8]^{\mathrm{t}}$$

encodes the situation where 8 judges chose the ranking $A_1A_2A_3$, 16 chose $A_1A_3A_2$, 6 chose $A_2A_1A_3$, and so on.

Given a profile **p**, a natural question to ask when comparing alternatives is whether the distribution P is the uniform distribution defined on $\mathcal{S}_m$. In other words, is it the case that $P[Y = y_i] = (1/m!)$ for all $y_i \in \mathcal{S}_m$? We will denote the null hypothesis that the distribution P is uniform on $\mathcal{S}_m$ by H_0.

Testing for differences among alternatives amounts to testing the null hypothesis H_0. To do so, a starting point to consider is the estimated *probabilities vector*

$$\hat{P} = \frac{1}{n}\mathbf{p}$$

that encodes the proportion of judges who choose each ranking in $\mathcal{S}_m$. If $\hat{P}$ is far from the constant vector $(1/m!)[1, \ldots, 1]^t$, then the judges would be favouring certain rankings and thus the null hypothesis H_0 of homogeneity would be rejected. Section 3 describes how this idea can be made precise and generalised, giving rise to many straightforward and useful tests of uniformity.

3. Linear rank tests of uniformity

Several nonparametric statistics exist to test the null hypothesis

$$H_0 : P = \text{Uniform } (\mathcal{S}_m).$$

Widely used tests include concordance tests, distance tests, and summary statistics tests (Cayley 1849; Friedman 1937; Kendall 1938; Kendall and Smith 1939; Wilcoxon 1945; Anderson 1959). A broad and important subclass of such tests are those whose test statistic is a function of the product $M\hat{P}$, where M is a $k \times m!$ matrix. Such tests are referred to as *linear rank tests*, the general construction of which we describe here.

Denote the *data space* $\mathbb{R}^{m!}$ by $\mathcal{D}$. Let $\mathcal{D}_0$ be the subspace of $\mathcal{D}$ that is orthogonal (with respect to the usual dot product) to the subspace spanned by the all-ones vector $[1, \ldots, 1]^t \in \mathcal{D}$. In other words, $\mathcal{D}_0$ is the subspace of vectors in $\mathcal{D}$ whose entries sum to zero.

Consider S, a nonzero subspace of $\mathcal{D}_0$. If $\mathbf{d} \in \mathcal{D}$, then we denote the projection of $\mathbf{d}$ onto S by $\mathbf{d}^S$. Thus, if $\hat{P}$ is an estimated probabilities vector in $\mathcal{D}$, then the closer $\hat{P}$ is to uniform, the closer the projection $\hat{P}^S$ is to the zero vector. As such, a natural statistic to consider for a test of uniformity is the length $\|\hat{P}^S\|$ of the projection of $\hat{P}$ onto S. The following is Theorem 3.1 in Marden (1995):

THEOREM 1   *Let S be a nonzero subspace of $\mathcal{D}_0$. If $\hat{P} = (1/n)\,\mathbf{p}$, and $\mathbf{p}$ is generated from a uniform distribution on $\mathcal{S}_m$ using n judges, then as $n \to \infty$, we have $nm!\|\hat{P}^S\|^2 \to \chi^2_{\dim(S)}$.*

Thus, if n is large enough, the null hypothesis H_0 is rejected when $nm!\|\hat{P}^S\|^2 > \chi^2_{\dim(S),\alpha}$ for a significance level α. We will refer to this test as the *linear rank test of uniformity associated with S*.

In practice, the subspace S is often associated with a specific summary statistic. In particular, suppose a summary statistic (e.g. mean ranks) can be computed via a linear transformation M defined on $\mathcal{D}$. We might then define S to be the projection of $\ker(M)^\perp$ (i.e. the orthogonal complement of the kernel of M) onto $\mathcal{D}_0$ and apply Theorem 1. A rejection of H_0 could then be explained using the computed summary statistic $M(\hat{P})$ because, by construction, $M(\hat{P})$ captures all of the information necessary to compute $\|\hat{P}^S\|^2$. For convenience, we will refer to $\ker(M)^\perp$ as the *effective space* of M (Daugherty, Eustis, Minton, and Orrison 2009).

Note that if the linear transformation M is encoded as a matrix (with respect to the usual basis of $\mathbb{R}^{m!}$), then $\ker(M)^\perp$ is simply the subspace of $\mathcal{D} = \mathbb{R}^{m!}$ spanned by the rows of M, and the

summary statistic is simply the matrix–vector product $M\hat{P}$. In general, however, computing the associated test statistic $nm!\|\hat{P}^S\|^2$ may be unwieldy when m is large because M and $\hat{P}$ are large. Fortunately, there are simple shortcuts for some popular summary statistics that allow one to compute $nm!\|\hat{P}^S\|^2$ directly from the entries in the vector $M\hat{P}$ (see Marden 1995, chap. 3). We describe three such popular summary statistics in Section 4.

4. Linear summary statistics

If asked to analyse rank data such as $\hat{P}$, a natural first step might be to compute some simple but potentially informative summary statistics. For example, the *marginals* summary statistic computes, for each alternative, the proportion of times that alternative is ranked first, second, third, and so on. The *means* summary statistic computes the average rank obtained by each alternative. The *pairs* summary statistic, on the other hand, computes for each ordered pair (A_i, A_j) of distinct alternatives, the proportion of judges who ranked A_i above A_j.

As noted in Section 3, well-known linear rank tests of uniformity are often associated with such summary statistics. For example, the Friedman test (Friedman 1937) uses the means summary statistic, and the Anderson test (Anderson 1959) uses the marginals summary. As such, we will refer to the Friedman test as the *means test*, and we will refer to the Anderson test as the *marginals test*. The means, marginals, and pairs tests are described in depth in Chapter 3 of Marden (1995).

The matrices needed to compute the means, marginals, and pairs summary statistics are relatively easy to construct. For example, if the rankings are listed lexicographically, $\hat{P} = \frac{1}{60}[6, 10, 6, 10, 14, 14]^t$, and

$$
M_{\mathrm{mrg}} =
\begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0
\end{bmatrix},
$$

then the entries in the product $M_{\mathrm{mrg}}\hat{P}$ are the marginal summary statistics:

$$
M_{\mathrm{mrg}}\hat{P} =
\begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0
\end{bmatrix}
\left(\frac{1}{60}
\begin{bmatrix}
6 \\ 10 \\ 6 \\ 10 \\ 14 \\ 14
\end{bmatrix}
\right)
=
\begin{bmatrix}
\frac{16}{60} \\ \frac{20}{60} \\ \frac{24}{60} \\ \frac{16}{60} \\ \frac{20}{60} \\ \frac{24}{60} \\ \frac{28}{60} \\ \frac{20}{60} \\ \frac{12}{60}
\end{bmatrix}
\begin{array}{l}
A \text{ ranked first} \\
A \text{ ranked second} \\
A \text{ ranked third} \\
B \text{ ranked first} \\
B \text{ ranked second} \\
B \text{ ranked third} \\
C \text{ ranked first} \\
C \text{ ranked second} \\
C \text{ ranked third}
\end{array}
$$

Similarly, for the means summary statistics, we can use the matrix

$$M_{\text{mns}} = \begin{bmatrix} 1 & 1 & 2 & 3 & 2 & 3 \\ 2 & 3 & 1 & 1 & 3 & 2 \\ 3 & 2 & 3 & 2 & 1 & 1 \end{bmatrix}$$

to compute the average rank of each alternative, and for the pairs summary statistic, we can use

$$M_{\text{prs}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

to compute the proportion of judges who ranked a given alternative above another.

5. Inconsistent test outcomes

In this section, we explain why different linear rank tests of uniformity can yield inconsistent test outcomes. Because of their popularity, we begin by considering the means, marginals, and pairs tests. Remarkably, if $m \geq 3$, then the effective spaces of the means, marginals, and pairs maps are related by a single orthogonal decomposition of the data space $\mathcal{D}$. The following theorem describes this decomposition.

THEOREM 2 *The data space* $\mathcal{D} = \mathbb{R}^{m!}$ *may be orthogonally decomposed as*

$$\mathcal{D} = W_1 \oplus W_2 \oplus W_3 \oplus W_4 \oplus W_5, \tag{2}$$

where

(1) W_1 *is the space spanned by the all-ones vector,*
(2) $W_1 \oplus W_2$ *is the effective space of the means matrix* M_{mns},
(3) $W_1 \oplus W_2 \oplus W_3$ *is the effective space of the marginals matrix* M_{mrg},
(4) $W_1 \oplus W_2 \oplus W_4$ *is the effective space of the pairs matrix* M_{prs}, *and*
(5) $\dim(W_1) = 1$, $\dim(W_2) = m - 1$, $\dim(W_3) = (m-1)(m-2)$, $\dim(W_4) = (m-1)(m-2)/2$, *and* $\dim(W_5) = m! - (3m^2 - 7m + 6)/2$.

Because the effective spaces for the means, marginals, and pairs summary statistics share some of the W_i, the results of one of the associated tests of uniformity can have implications for the other tests. To see why this is true, let $t_i = nm!\|\hat{P}^{W_i}\|^2$. Because the decomposition in Equation (2) is an orthogonal decomposition, it follows that

$$nm!\|\hat{P}\|^2 = t_1 + t_2 + t_3 + t_4 + t_5.$$

Furthermore, because $\mathcal{D}_0 = W_2 \oplus W_3 \oplus W_4 \oplus W_5$, the test statistic for the means test is simply t_2, for the marginals test it is $t_2 + t_3$, and for the pairs test it is $t_2 + t_4$. Thus, the t_i characterise those data vectors that lead to inconsistent results when using *any* combination of the means, marginals, and pairs tests of uniformity.

THEOREM 3 *Consider level α linear tests of uniformity, where $\gamma_{d,\alpha}$ is the critical value for d degrees of freedom.*

(1) *If the means and marginals tests disagree, then either $t_2 > \gamma_{m-1,\alpha}$ and $t_2 + t_3 \leq \gamma_{(m-1)^2,\alpha}$ or $t_2 \leq \gamma_{m-1,\alpha}$ and $t_2 + t_3 > \gamma_{(m-1)^2,\alpha}$.*
(2) *If the means and pairs tests disagree, then either $t_2 > \gamma_{m-1,\alpha}$ and $t_2 + t_4 \leq \gamma_{m(m-1)/2,\alpha}$ or $t_2 \leq \gamma_{m-1,\alpha}$ and $t_2 + t_4 > \gamma_{m(m-1)/2,\alpha}$.*
(3) *If the marginals and pairs tests disagree, then either $t_2 + t_3 > \gamma_{(m-1)^2,\alpha}$ and $t_2 + t_4 \leq \gamma_{m(m-1)/2,\alpha}$ or $t_2 + t_3 \leq \gamma_{(m-1)^2,\alpha}$ and $t_2 + t_4 > \gamma_{m(m-1)/2,\alpha}$.*

Given the decomposition in Equation (2), finding data vectors for which different tests of uniformity disagree is now relatively straightforward. To do so, find data vectors whose probabilities vectors satisfy at least one of the conditions on t_2, t_3, t_4 described in Theorem 3. As an example, let $m = 3$ and $\alpha = 0.05$, and consider the data vector

$$\mathbf{p} = \begin{bmatrix} 6 \\ 10 \\ 6 \\ 10 \\ 14 \\ 14 \end{bmatrix} \begin{array}{l} ABC \\ ACB \\ BAC \\ BCA \\ CAB \\ CBA \end{array}$$

for the three alternatives A, B, and C. Using the means test, the p-value is 0.0408, thus the null hypothesis is rejected. On the other hand, the p-values for the marginals and pairs tests are 0.1712 and 0.0937, respectively, which both fail to reject the null hypothesis when using either test.

To see why the inconsistencies in test results occur, decompose the profile as $\mathbf{p} = \mathbf{p}_1 + \mathbf{p}_2$, where $\mathbf{p}_i \in W_i$, and $\mathbf{p}_1 = [10, 10, 10, 10, 10, 10]^t$ and $\mathbf{p}_2 = [-4, 0, -4, 0, 4, 4]^t$. Thus, the data vector $\mathbf{p}$ is composed of vectors in just W_1 and W_2, which together form the effective space of the means summary statistic. In particular, the spaces W_3 and W_4 are not needed to construct $\mathbf{p}$. They are needed, however, to form the effective spaces of the marginals and pairs maps, which explains the larger p-values for the marginals and pairs tests.

Similarly, one can construct data vectors such that only the marginals uniformity test rejects the null hypothesis. For example, the data vector $\mathbf{p} = [8, 16, 6, 18, 10, 8]^t$ has p-values for the means, marginals, and pairs tests that are 0.8338, 0.0375, and 0.8232, respectively. Similarly, the data vector $\mathbf{p} = [15, 8, 7, 16, 17, 9]^t$ rejects the null hypothesis for the pairs test, but not for the means or marginals tests. The resulting p-values for the means, marginals, and pairs test are 0.8465, 0.9876, and 0.0396, respectively.

In order to further see why the inconsistencies occur, for each of the example data vectors, we find the projections into the subspaces W_1, W_2, W_3, W_4, and W_5. Table 1 provides an ANOVA-like display to illustrate the squared lengths of the projections into each of these spaces as well as R^2, the ratio of the squared length to the sum of the squared lengths without considering the constant space W_1 (see, Marden 1995, for similar tables). As one can see from the table, the data vector $[6, 10, 6, 10, 14, 14]$ has a large projection into W_2 thus causing the means test to reject, the data vector $[8, 16, 6, 18, 10, 8]$ has a large projection into W_3 leading the marginals test to reject, and the data vector $[15, 8, 7, 16, 17, 9]$ has a large projection into W_4 corresponding to the pairs test rejecting.

The decomposition of the data space given in Theorem 2 explains how different data structures affect the outcomes of the means, marginals, and pairs tests. Furthermore, Theorem 3 provides the exact conditions a data profile must meet to lead to inconsistent results for these tests.

In general, given two distinct subspaces S and S' of $\mathcal{D}_0$, there will always exist profiles for which the linear rank tests of uniformity associated with S and S' will disagree.

Table 1. Data vector projections.

Subspace	Dim	[6, 10, 6, 10, 14, 14]		[8, 16, 6, 18, 10, 8]		[15, 8, 7, 16, 17, 9]	
		SS	100*R^2	SS	100*R^2	SS	100*R^2
W_1	1	600		726		864	
W_2	2	64	1	4	0.0339	4	0.04
W_3	2	0	0	108	0.9153	0	0
W_4	1	0	0	6	0.0508	96	0.96

THEOREM 4 *Let S and S′ be nonzero subspaces of $\mathcal{D}_0$, and let $0 < \alpha < 1$. If $S \neq S′$, then there exist profiles $\mathbf{p} \in \mathcal{D}$ such that the level α test of uniformity for $\hat{P}$ associated with S will reject the null hypothesis H_0, while the test associated with S′ will fail to reject H_0.*

There do exist, however, profiles for which all possible linear rank tests of uniformity will agree. The following theorem and corollary highlight this point.

THEOREM 5 *Let $\hat{P}$ be a sample probabilities vector when n judges are asked to rank m alternatives. Every level α linear rank test of uniformity associated with a d-dimensional subspace of $\mathcal{D}_0$ will fail to reject H_0 if and only if $nm!\|\hat{P}^{\mathcal{D}_0}\|^2 < \gamma_{d,\alpha}$.*

Because $\gamma_{d,\alpha} \leq \gamma_{d′,\alpha}$ whenever $d \leq d′$, the following corollary provides a condition that the data profile must satisfy to ensure that all possible linear rank tests of uniformity will reject the null hypothesis.

COROLLARY 6 *If $nm!\|\hat{P}^{\mathcal{D}_0}\|^2 < \gamma_{1,\alpha}$, then every level α linear rank test of uniformity will fail to reject H_0.*

It is not, however, possible to find a profile for which *all* linear rank tests of uniformity will reject the null hypothesis.

THEOREM 7 *Let $m \geq 3$, and let $\mathbf{p} \in \mathcal{D}$ be a profile. Then there exists a subspace S of $\mathcal{D}_0$ such that the associated linear rank test of uniformity will fail to reject the null hypothesis.*

The results in this section highlight how linear rank tests of uniformity are related to their associated effective spaces. Viewing the tests in this framework brings to light relationships between different tests. Using only the effective spaces of tests, data profiles for which tests will yield inconsistent results can be characterised, thus making it easy to construct examples of profiles that cause disagreements.

6. Power

The most powerful linear rank tests of uniformity will be those whose associated subspaces S have the property that the underlying probability distribution P (when viewed as a vector in $\mathcal{D}$) is contained in the subspace of $\mathcal{D}$ spanned by S and the all-ones vector. This is because the linear rank test of uniformity associated with S will detect a deviation from uniformity only if the data profile $\mathbf{p}$ generated by P is largely contained in S. The power of a test relative to a nonuniform distribution is therefore related to whether the distribution P will yield profiles with large projections into the test's effective space. To illustrate this, Table 2 outlines the power results for three example non-null distributions for the $m = 3$ case.

Table 2.   Power table.

Non-null distribution	Power
$[\frac{6}{60}, \frac{10}{60}, \frac{6}{60}, \frac{10}{60}, \frac{14}{60}, \frac{14}{60}]$	Means: 0.995 Marginals: 0.986 Pairs: 0.992
$[\frac{8}{60}, \frac{14}{60}, \frac{8}{60}, \frac{8}{60}, \frac{14}{60}, \frac{8}{60}]$	Means: 0.061 Marginals: 0.917 Pairs: 0.054
$[\frac{13}{60}, \frac{7}{60}, \frac{7}{60}, \frac{13}{60}, \frac{7}{60}, \frac{13}{60}]$	Means: 0.047 Marginals: 0.06 Pairs: 0.956

Column 1 of the table specifies a non-null distribution and column 2 displays the power results of the means, marginals, and pairs test for 200 judges and 1000 simulations. The non-null distributions listed in Table 2 were purposely constructed to be completely contained in $W_1 \oplus W_2$, $W_1 \oplus W_3$, and $W_1 \oplus W_4$, respectively.

For data in $W_1 \oplus W_2$ ($[\frac{6}{60}, \frac{10}{60}, \frac{6}{60}, \frac{10}{60}, \frac{14}{60}, \frac{14}{60}]$), the means, marginals, and the pairs all yield high power. This is because W_2 is contained in the effective space of the matrices associated with each of the tests. On the other hand, data contained only in $W_1 \oplus W_3$ ($[\frac{8}{60}, \frac{14}{60}, \frac{8}{60}, \frac{8}{60}, \frac{14}{60}, \frac{8}{60}]$) will produce high power for the marginals test but not the others because W_3 is contained only in the effective space of the marginals matrix. Because W_4 is only contained in the pairs effective space, we see high power for the pairs test but not the others when data are generated from a non-null distribution completely contained in $W_1 \oplus W_4$ ($[\frac{13}{60}, \frac{7}{60}, \frac{7}{60}, \frac{13}{60}, \frac{7}{60}, \frac{13}{60}]$).

The power results in Table 2 highlight the fact that when a non-null distribution produces data whose projection into the effective space of a particular test is large, then that test has high power. On the other hand, if the data substantially misses the effective space of a particular test, then that test will have low power.

7. W_5 and the probabilities test

By Theorem 2, the data space $\mathcal{D}$ can be decomposed as

$$\mathcal{D} = W_1 \oplus W_2 \oplus W_3 \oplus W_4 \oplus W_5,$$

where the subspaces W_1, W_2, W_3, and W_4 are directly related to the means, marginals, and pairs tests. No vector in the subspace W_5, however, is used by any of these popular tests. Thus, if $\hat{P}$ has a large projection into W_5, the means, marginals, and pairs could fail to reject the null hypothesis even though nonuniformity may be heavily present. This could be an issue because the dimension of W_5 eventually dwarfs the dimension of $W_1 \oplus W_2 \oplus W_3 \oplus W_4$.

THEOREM 8   *If $m = 3$, then* $\dim(W_5) = 0$. *However, as $m \to \infty$, then* $\dim(W_5)/\dim(\mathcal{D}) \to 1$.

To illustrate how rapidly W_5 grows, Table 3 contains the values of the ratio $\dim(W_5)/\dim(\mathcal{D})$ for $m = 3$ to $m = 10$ alternatives. As the table shows, even when $m = 4$, the means, marginals, and pairs tests could easily fail to detect nonuniformity for many profiles.

Table 3. The ratio $\dim(W_5)/\dim(\mathcal{D})$.

m	$\frac{\dim(W_5)}{\dim(\mathcal{D})}$
3	0.000
4	0.458
5	0.808
6	0.950
7	0.990
8	0.998
9	0.999
10	1.000

As an example of such a profile when $m = 4$, consider the following data profile, where the rankings are listed in a lexicographical order:

$$\mathbf{p} = [30, 10, 10, 30, 30, 10, 10, 30, 30, 10, 10, 30, 30, 10, 10, 30, 30, 10, 10, 30, 30, 10, 10, 30]^\text{t}.$$

This profile is completely contained in $W_1 \oplus W_5$. Thus, the p-values for the means, marginals, and pairs tests are all approximately 1, even though $\mathbf{p}$ appears to be the result of a highly nonuniform distribution P.

One linear rank test that does take advantage of W_5 is the *probabilities test* (Marden 1995, chap. 3). This test considers the proportion of judges that choose each different ranking. Its effective space is all of $\mathcal{D}_0 = W_2 \oplus W_3 \oplus W_4 \oplus W_5$. The associated linear transformation M is simply the identity map: $M\hat{P} = \hat{P}$.

The probabilities test is straightforward to use, but for any underlying distribution P, it will always be possible to construct a test that is more powerful than the probabilities test. Furthermore, the number of judges n may not be big enough for the limiting distribution of the test statistic to be close to χ^2. Ideally, we would like to use a test whose associated subspace S captures most of P, but whose dimension is not too large. The construction of new tests is addressed in Sections 8 and 9.

8. Constructing new tests

The relatively large dimension of W_5 suggests that it would be useful to be able to construct new tests.

Theorem 1 showed that the distribution of the test statistic $nm!\|\hat{P}^S\|^2$ is approximately χ^2 with degrees of freedom equal to the dimension of S. The question arises, however, about how large n must be in order for the approximation to be accurate. The size of n will limit the tests that are available. On the other hand, in the case where n is small, it might be worthwhile to use simulations to estimate a p-value. Also, as Marden (1995, p. 93) suggests, the size of n may provide guidance for which test to use:

> I would tend to concentrate on the four general test statistics, using whichever the data can bear: If $n \geq 5\,m!$, the Probabilities test; if $n \geq 5\,m^2$, the Marginals test; if $n \geq 3\,m^2$, the Pairs test, and if $n \geq 3\,m$, the Means test. These recommendations are very conservative, based on having about five observations per degree of freedom.

The difference between the value $5\,m!$ and the values $5\,m^2$, $3\,m^2$, and $3\,m$ is sizable, even for small m. So another reason for wanting to construct a new test is when you have a data set where the number of judges is more than $5\,m^2$ but much less than 5 m!.

There may also be a specific type of nonuniformity that is captured particularly well by a specific subspace S. In this case, one could conservatively plan on sampling $5(\dim S)$ judges to

collect your data profile. For example, suppose $m = 3$ and that you believe the nonuniformity of P will easily be captured by focusing on the number of times an alternative is ranked in first place. One could then let S be the projection into $\mathcal{D}_0$ of the subspace spanned by the three vectors

$$
\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad
\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}
\quad
\begin{matrix} ABC \\ ACB \\ BAC \\ BCA \\ CAB \\ CBA \end{matrix}
$$

The first vector corresponds to judges ranking alternative A in first place, the second vector does the same for alternative B, and the third for alternative C. Therefore, the subspace spanned by these three vectors will capture nonuniformity among the number of times each alternative is ranked in first place.

Finally, when constructing a linear rank test of uniformity, it seems natural to require that it be *label independent*. In other words, it should not be the case that swapping any of the labels of the alternatives might reverse the outcome of the test. For example, if $\mathbf{p} = [8, 16, 6, 18, 10, 8]^t$ and we swap the labels for A and B, then the resulting profile would be $\mathbf{p}' = [6, 18, 8, 16, 8, 10]^t$. As such, we would want to use only subspaces S of $\mathcal{D}_0$ with the property that $\|\mathbf{p}^S\| = \|\mathbf{p}'^S\|$. The means, marginals, pairs, and probabilities tests, as well as the first place test suggested in this section are all label-independent.

It should be noted that other subspaces of the data space $\mathcal{D}$ that grow slowly in dimension have been well studied. More specifically, Diaconis has written about subspaces of $\mathcal{D}$ that naturally generalise the effective space of the marginals summary statistic. Examples include subspaces that capture the number of times a subset (of two or more) alternatives has occupied a particular subset of positions in the rankings. See Diaconis (1988, 1989) for details concerning the resulting *spectral decomposition* of $\mathcal{D}$.

Another well-studied decomposition of $\mathcal{D}$ is often referred to as the *inversion* decomposition, which is discussed in Grossman and Minton (2009) and McCullagh (1993). The associated subspaces in this case naturally generalise the effective space of the pairs summary statistic, and are based on summary statistics concerning triples of alternatives, quadruples of alternatives, and so on. For an overview and comparison of the spectral decomposition and the inversion decomposition, see Marden (1995, chap. 2).

In general, there are two issues to consider when discussing the choice of a linear test of uniformity. The first issue is to make sure the dimension of the subspace relates well to the number of observations one has available. The second issue is to determine the type of nonuniformity that is interesting to uncover. The subspaces in the spectral and inversion decompositions grow relatively slowly and thus do not have the large gaps in dimensionality like the decomposition on which we have been focusing. However, the motivation for using the spectral and inversion decompositions (which is based on an understanding of the type of nonuniformity the decompositions uncover) currently seems to be lacking. Thus, this paper focuses on the decomposition described because the use of the associated tests is so prevalent.

In summary, to construct a new test, one should consider the number of observations available in the data set to ensure accurate approximation of the limiting distribution of the test statistic. A very conservative test will have at most $n/5$ degrees of freedom. In addition, one must define a set of functions that will define and span the space S. In order to choose these functions, it may prove useful to consider the type of nonuniformity that is believed to exist in the data and one is wanting to uncover. The following section illustrates the construction of a new test using a real data example.

9. Example

To illustrate the ideas presented in the previous sections, we use ranking data collected in 2006 at the College of Education at the University of Missouri (Rohs 2007). A set of 58 inservice teachers were administered the Early Childhood Belief Survey aimed at gauging early childhood teacher beliefs about what influences teacher lesson planning and instruction. Teachers were asked to provide full rankings of seven items from least influential to most influential on their lesson planning and their instruction. The items were parents, school system policy, principal, teacher, state regulations, other teachers, and school advisory council.

The researcher administering this survey was interested in the influence on curriculum planning by people and policies. The specific subset of items of interest were parents, A; school system policy, B; teacher, C; and state regulations, D . These items naturally split into two categories, individual items and policy items. An interesting test of uniformity to construct would thus be one that might determine whether teachers recognise the split. In general, we answer: do teachers believe the four items influence curriculum planning in different ways?

To answer the research question, the full seven alternative data set is reduced and re-ranked to only include the four items of interest. There are 24 possible ranking outcomes which create a profile $\mathbf{p} \in \mathbb{R}^{24}$. The data are summarised in Table 4.

To construct a test of uniformity, we consider the number of observations available in the data set as well as the specific interest of the researcher. Based on the relatively small number of observations in the data set (58), the data can only bear a test that is at most 12 or 13 dimensional. Because the researcher is interested in picking up differences among four items that group into two natural categories, we construct a test that determines whether teachers rank pairs of items together. As such, we define functions that pick up whether the alternatives are split into two subsets in all possible ways and ranked in all possible ways. There are thus 18 natural functions to consider. An example function with the alternatives listed lexicographically is

$$[1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0].$$

This function counts the number of times the pair of alternatives A (parents) and B (school system policy) are ranked simultaneously in the first and second spots. The first two ones in the vector represent when A is ranked first and B is ranked second, while the second set of two ones captures when A is ranked second and B is ranked first. Another function to consider would be:

$$[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0].$$

In this case, this function still considers the pair A and B, but it can be used to count the number of times A and B are ranked in the second and third position. In this sense, the space *S* spanned by the 18 vectors of this type captures all of the possible ways to partition the alternatives into two

Table 4. Inservice teacher data.

No. of teachers	Ranking	No. of teachers	Ranking	No. of teachers	Ranking
1	1234	7	2314	1	3412
0	1243	0	2341	1	3421
7	1324	1	2413	0	4123
3	1342	2	2431	3	4132
6	1423	2	3124	1	4213
0	1432	1	3142	8	4231
1	2134	3	3214	4	4312
1	2143	4	3241	1	4321

groups of two. The linear rank test of uniformity associated with this subspace *S* yields the *p*-value 0.00002 and thus rejects the null. This test therefore detects nonuniformity present in these data. In particular, the nonuniformity can be described by teachers grouping sets of two items together.

Knowing that teachers believe that these alternatives are significantly affecting curriculum planning in different ways, a researcher may use this as a springboard to unpack the data set and perform *post hoc* analysis to uncover specific relationships. This analysis is similar to the analysis done in Diaconis (1989). For example, Tables 5 and 6 provide the joint distributions of the individual and the policy alternatives.

From the joint distributions, we see that it is common for teachers to either rank both individual items together or both policy items together. For example, there are $21(13 + 8)$ teachers who rank parents and teachers as their top two alternatives and another $16(11 + 5)$ teachers who rank them as their bottom two alternatives. Thus, a total of 37 teachers keep the individual items together as a pair – either ranking them high or low. These same 37 teachers rank the policy items together as well. This means that not only do teachers split the items into groups of two, but they split according to the two natural sets. This suggests a polarisation among the teachers – individual driven and policy driven. The researcher viewed the two sets as essentially being opposite. Individual items are those closest to a teacher, while policy items are the most global and farthest from the teacher's nucleus. The main scope of the researcher in the College of Education was to determine the most effective ways to affect curricular change. The results here suggest that change may be implemented through policy for some teachers and through individual support for others. A next step for this researcher would then be to determine if there are factors that predict a teacher type.

Using these same data, we also ran the means, marginals, pairs, and probabilities tests. These tests yield *p*-values of 0.7979, 0.1312, 0.6468, and 0.0002, respectively. Table 7 illustrates the projection breakdown into each of the subspaces of the decomposition. Interestingly, only the probabilities test picks up nonuniformity. However, for these data, the probabilities test has 23 degrees of freedom and thus one must question the appropriateness of its use in this scenario. Although the other three tests are usable with degrees of freedom 3, 9, and 5, respectively, they do not reject the null and fail to pick up nonuniformity. It is clear from looking at the joint distribution tables, however, that nonuniformity is present.

Table 5. Joint distribution of parents and teacher.

Parents	Teacher			
	1	2	3	4
1	0	13	1	3
2	8	0	3	1
3	4	3	0	5
4	5	1	11	0

Table 6. Joint distribution of school system policy and state regulations.

School	State			
	1	2	3	4
1	0	4	1	3
2	12	0	1	4
3	1	7	0	14
4	3	1	7	0

Table 7. Example data vector projection.

Subspace	Dim	Vector	
		SS	$100*R^2$
W_1	1	140.1667	
W_2	3	2.4500	0.0183
W_3	6	30.8000	0.2301
W_4	3	7.7500	0.0579
W_5	11	92.8333	0.6936

This section thus provides a concrete example of how to construct a new test. In addition, the construction process highlights how to utilise the researcher's knowledge about the nature of the items to uncover nonuniformity. For the real data example presented in this section, if one used the means, marginals, or the pairs test, the nonuniformity would not have been picked up.

10. Conclusion

Several types of nonparametric tests exist to analyse ranked data of the form of repeated measures for a set of *m* alternatives. The problem of *n* rankings asks whether or not *m* fully ranked alternatives by a sample of *n* judges are significantly different. In order to answer this question, the null hypothesis of uniformity over the set of permutations of the *m* alternatives is tested. A popular class of such tests is *linear rank tests of uniformity*. These tests are often associated with common summary statistics making them particularly natural options for testing the null hypothesis. However, the results in this paper highlight that different tests can lead to different results. In addition, we show that the current existing tests may actually not be feasible for examining specific data sets and data structures.

By decomposing the ranked data space, we show that all of these discrepancies among test outcomes are due to the manner in which each test uses the data. If the data are largely present in a test's associated effective space, then the test will be able to pick out nonuniformity. However, in the case where the data lie in other parts of the data space, the test will not capture the nonuniformity. In addition, our results make clear the relationship between the effective space, the test outcome, and the power of a test. This work therefore makes explicit a conceptual framework to better understand linear rank tests of uniformity discussed in Marden (1995).

Finally, we gave a concrete example illustrating the process of constructing a new test. In particular, to ensure an accurate approximation of the limiting distribution of the test statistic, one should consider the dimension of *S*. Ultimately, constructing a new test amounts to defining a set of functions that span a desired subspace. The choice of functions may rely on the type of nonuniformity the researcher believes that may exists in the data. As explained in the previous section, the construction of new linear rank tests of uniformity is straightforward and worthwhile.

11. Proofs

Proof of Theorem 2   The statements in Theorem 2 follow directly from the discussions found in Sections 6 and 7 of Daugherty et al. (2009), Section 8B in Diaconis (1988), and Section 2.6.1 in Marden (1995). In particular, Marden (1995, eq. (2.78)) shows that the data space $\mathcal{D}$ can be decomposed orthogonally as

$$\mathcal{D} = \bigoplus_{\lambda} V^{\lambda},$$

where the direct sum is over all partitions of m, and the V^λ (to use the notation found in Marden (1995)) form the *canonical* decomposition of $\mathcal{D}$ (see Theorem 1 in Section 8B of Diaconis (1988)). Therefore, in Theorem 2, $V^{(m)} = W_1$, $V^{(m-1,1)} = W_2 \oplus W_3$, and $V^{(m-2,1,1)} \supset W_4$. Furthermore, if we denote the effective space of a matrix M by $E(M)$, then $W_2 = E(M_{\text{mns}}) \cap V^{(m-1,1)}$, $W_3 = W_2^\perp \cap V^{(m-1,1)}$, and $W_4 = E(M_{\text{prs}}) \cap V^{(m-2,1,1)}$. (See, for example, Theorem 6 in Daugherty et al. (2009)). ∎

Proof of Theorem 3   In order for two selected tests to disagree, one test must have a statistic larger than the critical value while the other must have a statistic smaller than the critical value for a fixed significance level α and their associated degrees of freedom. Letting $t_i = nm! \|\hat{P}^{W_i}\|^2$ and by Theorem 2, the statistics for the means, marginals, and pairs tests are t_2, $t_2 + t_3$, and $t_2 + t_4$, respectively. Thus, by comparing the means and the marginals test outcomes in (1), disagreement will occur when either t_2 is larger than the critical value $\gamma_{m-1,\alpha}$ and $t_2 + t_3$ is smaller than the critical value $\gamma_{(m-1)^2,\alpha}$ or vice versa. Statements (2) and (3) follow in a similar manner. ∎

Proof of Theorem 4   If S is not a subspace of S', then there exist vectors in S that are not in S'. Thus, for any $a, b \in \mathbb{R}$ such that $0 < a < b$, there exist profiles **p** such that $\|\mathbf{p}^S\| > b$ and $\|\mathbf{p}^{S'}\| < a$. For these profiles, if b is large enough and a is small enough, then the test associated with S will reject H_0 but the test associated with S' will not.

On the other hand, suppose S is a subspace of S'. Then for any profile **p**, we have that $\mathbf{p}^S \in S'$. Because $S \neq S'$, however, we know that $\dim S < \dim S'$. Thus, the critical value associated with S is less than the critical value associated with S'. In this case, we simply choose profiles **p** so that

$$\gamma_{\dim S, \alpha} < \|\mathbf{p}^S\| < \gamma_{\dim S', \alpha}.$$

As before, for these profiles the test associated with S will reject H_0 but the test associated with S' will not. ∎

Proof of Theorem 5   This follows directly from the fact that, for all subspaces S of $\mathcal{D}_0$, $\|\hat{P}^S\| \leq \|\hat{P}^{\mathcal{D}_0}\|$. ∎

Proof of Theorem 7   Let S be the orthogonal complement in $\mathcal{D}_0$ of the space spanned by $\mathbf{p}^{\mathcal{D}_0}$. Then $\|\mathbf{p}^S\| = 0$, and the test associated with S will therefore fail to reject H_0. ∎

Proof of Theorem 8   By Theorem 1, the dimension of W_5 is $m! - (3m^2 - 7m + 6)/2$, thus $\dim(W_5)/\dim(\mathcal{D}) = m! - (3m^2 - 7m + 6)/2/m!$ which as $m \to \infty$ clearly equals 1. ∎

References

Anderson, R.L. (1959), 'Use of Contingency Tables in the Analysis of Consumer Preference Studies', *Biometrics*, 15, 582–590.

Cayley, A. (1849), 'A Note on the Theory of Permutations', *Philosophical Magazine*, 34, 527–529.

Daugherty, Z., Eustis, A., Minton, G., and Orrison, M. (2009), 'Voting Theory, the Symmetric Group, and Representation Theory', *The American Mathematical Monthly*, 116(8), 667–687.

Diaconis, P. (1988), *Group Representations in Probability and Statistics*, Hayward, CA: Institute of Mathematical Statistics.

Diaconis, P. (1989), 'The 1987 Wald Memorial Lectures: A Generalization of Spectral Analysis with Application to Ranked Data', *The Annals of Statistics*, 17(3), 949–979.

Friedman, M. (1937), 'The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance', *Journal of the American Statistical Association*, 32, 675–701.

Grossman, J.P., and Minton, G. (2009), 'Inversions in Ranking Data', *Discrete Mathematics*, 20(309), 6149–6151.

Hajek, J., Sidak, Z., and Sen, P. (1999), *Theory of Rank Tests*, London: Academic Press.

Kendall, M.G. (1938), 'A New Measure of Rank Correlation', *Biometrika*, 30, 81–93.

Kendall, M.G., and Smith, B.B. (1939), 'The Problem of *m* rankings', *American Mathematical Society*, 10, 275–287.

Marden, J.I. (1995), *Analyzing Ranked Data*, London: Chapman & Hall.

McCullagh, P. (1993), 'Permutations and Regression Models', in *Probability Models and Statistical Analyses for Ranking Data*, eds. M.A. Fligner and J.S. Verducci, New York: Springer, pp. 196–215.

Rohs, J. (2007), 'Child Outcomes in Head Start Classrooms: The Impact of Teacher Beliefs and Interactions', unpublished doctoral dissertation, University of Missouri-Kansas City.

Spearman, C. (1904), 'The Proof and Measurement of Association Between Two Things', *American Journal of Psychology*, 15, 72–101.

Wilcoxon, F. (1945), 'Individual Comparisons by Ranking Methods', *Biometrics Bulletin*, 1, 80–83.