the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2022), 75, 334–362
© 2022 The Authors. *British Journal of Mathematical and Statistical Psychology* published by John Wiley & Sons Ltd on behalf of British Psychological Society

www.wileyonlinelibrary.com

Remarkable properties for diagnostics and inference of ranking data modelling

Cristina Mollica and Luca Tardella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy

The Plackett-Luce model (PL) for ranked data assumes the forward order of the ranking process. This hypothesis postulates that the ranking process of the items is carried out by sequentially assigning the positions from the top (most liked) to the bottom (least liked) alternative. This assumption has been recently relaxed with the Extended Plackett-Luce model (EPL) through the introduction of the discrete reference order parameter, describing the rank attribution path. By starting from two formal properties of the EPL, the former related to the inverse ordering of the item probabilities at the first and last stage of the ranking process and the latter well-known as independence of irrelevant alternatives (or Luce's choice axiom), we derive novel diagnostic tools for testing the appropriateness of the EPL assumption as the actual sampling distribution of the observed rankings. These diagnostic tools can help uncovering possible idiosyncratic paths in the sequential choice process. Besides contributing to fill the gap of goodness-of-fit methods for the family of multistage models, we also show how one of the two statistics can be conveniently exploited to construct a heuristic method, that surrogates the maximum likelihood approach for inferring the underlying reference order parameter. The relative performance of the proposals, compared with more conventional approaches, is illustrated by means of extensive simulation studies.

1. Introduction

Psychological and behavioural studies typically investigate personality traits, such as preferences and attitudes, that cannot be directly observed or are difficult to measure. For this reason, research in these fields is often conducted by assessing choice and decision processes that lead to the collection of ordinal data, rather than observations on the numerical scale.

Let us consider, for example, an experiment in which a sample of N judges are asked to rank a set $I = \{1,\ldots,K\}$ of K labelled alternatives, namely *items*, according to a certain criterion. The final outcome of the comparative evaluation is an ordered sequence collecting the positions attributed to each object, called ranking. Formally, a *ranking* is a vector $\pi = (\pi(1),\ldots,\pi(K))$ where the entry $\pi(i)$ indicates the position attributed to the ith alternative. Equivalently, data can be recorded in the *ordering* format $\pi^{-1} = (\pi^{-1}(1),\ldots,\pi^{-1}(K))$, where the generic component $\pi^{-1}(j)$ indicates the item ranked in the jth

Correspondence should be addressed to Cristina Mollica, Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Roma, Italy (email: cristina.mollica@uniroma1.it).

DOI:10.1111/bmsp.12260

position. This implies that ranking/ordering data are multivariate ordinal data taking values in the set of permutations S_K of the first K integers.

These observations are common also in other areas of research, involving market surveys on customers' preferences for consumer goods (Gormley & Murphy, 2010; Yao & Böckenholt, 1999) and voting systems allowing for the elicitation of an ordering of the candidates in elections (Gormley & Murphy, 2008; Lee & Yu, 2012). Ranking data emerge also from the literature on expert opinion elicitation when, in order to reduce the elicitation burden, judgements are expressed in ordinal form instead of as numerical assessments (Overland & Juraev, 2019; Wang & Bier, 2013). Another relevant field concerns the competition/sport context, where the competitors (players, teams) are ranked according to a certain measure of ability, such as the finishing time in a race or the score amassed during a championship (Henderson & Kirrane, 2018; Henery, 1981; Stern, 1990).

The broad statistical literature on methods and models for analysing ranking data is reviewed in Marden (1995) and, more recently, in Alvo and Yu (2014) and Liu, Crispino, Scheel, Vitelli, and Frigessi (2019). By focusing on the parametric modelling approach, distributions for random permutations are traditionally classified into four main categories: (i) order statistics models (OSMs), whose seminal work is represented by Thurstone (1927); (ii) paired comparison models (Bradley, 1976, 1984); (iii) distance-based models (DBMs) (Fligner & Verducci, 1986; Mallows, 1957); and (iv) stagewise models (Fligner & Verducci, 1988). This work concentrates on the parametric family of type (iv), relying on the idea that the ranking process can be decomposed into consecutive stages for each position that has to be assigned, in particular on the *Extended Plackett–Luce* model (EPL) introduced by Mollica and Tardella (2014). The EPL generalizes the popular *Plackett–Luce* model (PL), presented by Luce (1959) and Plackett (1968), by relaxing the implicit *forward order* assumption, according to which the ranking process of the alternatives proceeds sequentially from the most liked to the least liked item. This extension was accomplished by adding the *reference order parameter* $\rho = (\rho(1),\ldots,\rho(K))$ in the PL formulation. It indicates the rank assignment order, that is, the component $\rho(t)$ denotes the position attributed at the tth stage.

One aspect which is very often overlooked in the ranking data literature concerns the assessment of model adequacy for the observed data. Traditional approaches are based on the construction of diagnostics to detect possible lack of fit of generic sample quantities, rather than assessing the conformity of the data with peculiar features of the postulated model. Moreover, the investigation of the effectiveness of these methods has been limited to a few parametric families, such as the DBM (Cohen & Mallows, 1983; Feigin & Cohen, 1978) and the OSM (Tsai & Yao, 2000; Yao & Böckenholt, 1999; Yu, 2000). Nevertheless, model-specific diagnostics are valuable to support the crucial phase of model building, that is, to motivate the adoption of a certain parametric distribution with the aim of optimizing data description or gaining computational convenience.

These arguments motivated us to review the existing methods and develop some original tools to appropriately check the model misspecification issue for the class of multistage models, specifically for the EPL assumption as the data-generating mechanism. We first introduced two novel test statistics: the former is based on a formal property of the EPL class which, to the best of our knowledge, has not been highlighted earlier in the literature, while the latter relies on the well-known assumption of the PL distribution, known as the *independence of irrelevant alternatives* or *Luce's choice axiom* (Luce, 1959). We believe that the former property can be of great interest in those contexts where some idiosyncratic attitude could guide the ranker to assign positions in a noncanonical way, deviating from the most common and natural forward or backward orders. Some evidence of this phenomenon has been highlighted in some real data applications (Mollica & Tardella, 2021). Through an extensive simulation study under different model specifications, we compared the power of the two proposed diagnostics with that of more frequently used test statistics. The analysis of simulated data revealed the relative merits and limits of the competing diagnostics for the EPL assumption, allowing us to provide some reasonable general guidelines.

As a by-product of our interest in the stagewise parametric class, we further considered the former EPL property from the inferential perspective as the key element of a heuristic method to estimate ρ. We implemented a simulation study to quantify the inferential ability of the proposed likelihood-free estimation strategy to recover the actual reference order parameter. It showed promising and consistent behaviour by the heuristic technique, which could be effectively exploited to reduce the computation burden affecting the EPL estimation task as well as for the relative assessment of model goodness of fit.

This paper is organized as follows. In Section 2 a review of the fundamental ranking models is provided, with a special focus on the EPL and related inferential approaches. Section 3 gives a detailed overview of the strategies proposed in the literature to address the model assessment issue for ranking distributions within the frequentist and Bayesian paradigms. Two novel goodness-of-fit diagnostics for the EPL parametric class are then introduced in Section 4, and a comparative evaluation with more standard measures of model adequacy follows in Section 5. The application of the model assessment tools considered is illustrated in Section 6. An original heuristic method to infer the reference order parameter from the likelihood-free perspective is defined in Section 7, where its effectiveness is also investigated with an extensive simulation study. Concluding remarks and proposals for future work are discussed in Section 8.

2. An overview of parametric ranking models

2.1. Ordered statistics models

The class of OSMs, also known as *random utility models*, was originally introduced in Thurstone (1927). Thurstone proposed the existence of a unobserved quantitative mechanism underlying the ranking process, such that each item i is associated with a continuous latent random variable (r.v.) W_i, also called *score* or *utility*. The score should be intended as a latent item feature, measurable on a unidimensional scale and on which the comparative judgement is based, such as a preference/liking measure. In this perspective, the *Thurstone model* (TM) can be defined by assuming a parametric joint distribution for the Ws and considering the ranking sequence as the result of the ordering of the item utilities with probability given by

$$\boldsymbol{P}(\pi) = \boldsymbol{P}(W_{\pi^{-1}(1)} < \ldots < W_{\pi^{-1}(K)}) \qquad \pi \in \mathcal{S}_K. \tag{1}$$

Originally, the TM involved only two items, and Daniels (1950) extended it for $K > 2$ alternatives.

By postulating that the Ws are independent and normally distributed with different means but equal variances, expression (1) translates into the *Thurstone–Mosteller–Daniels* model or *Case V* model (Daniels, 1950; Mosteller, 1951). More general versions of this approach relax the hypothesis of homoscedasticity of the independent normal distributions (*Case III* model) or contemplate other parametric laws for the item scores. For example, Henery (1983) and Luce (1959) considered independent Gamma and Gumbel distributions, respectively. The Gumbel case is very popular because it leads to the PL.

The flexibility in the choice of the latent multivariate model makes the OSM class a very broad family of ranking distributions. However, the most common specifications postulate independent utilities whose distribution belongs to a location family with possibly different location parameters. Under this assumption, the TM satisfies a specific property not shared by the other classes, which we will later consider to clarify the relationship with the EPL, such that the relative order of any subset of the items is independent of the ordering of any disjoint subset (Critchlow, Fligner, & Verducci, 1991).

2.2. Ranking models based on paired comparisons

This approach relies on the possibility of converting a ranked sequence into the corresponding set of $K(K-1)/2$ comparisons between pairs of items. Nevertheless, a generic set of paired comparisons (PCs) is not necessarily consistent with the definition of a ranking. In fact, intransitivities (or circularities) of the type $\pi(1) < \pi(2) < \pi(3) < \pi(1)$ are allowed in pairwise comparison modelling, whereas they are not consistent with a ranking elicitation.

Let us define with $X_{ii'} \sim Bern(\eta_{ii'})$ the distribution of the binary r.v. indicating the preference for item i over i' in the paired comparison ($1 \leq i < i' \leq K$). By assuming that all the $K(K-1)/2$ comparisons are drawn independently and governed by the probabilities $\eta_{ii'}$, the sequence of paired preferences is considered valid if it does not contain any circularity, otherwise it is discarded and the comparisons are repeated until no circularity is present. In this setting, the probability of each ranking is

$$\boldsymbol{P}(\pi|\underline{\eta}) \propto \prod_{i<i'} \eta_{ii'}^{x_{ii'}(\pi)} (1-\eta_{ii'})^{1-x_{ii'}(\pi)} \qquad \pi \in S_K,$$

which is known in the literature as the *Babington Smith model* (BSM), originally proposed by Babington Smith (1950). By setting special forms for the Bernoulli probabilities, popular subclasses of the BSM can be derived. For example, Bradley and Terry (1952) introduced parameters $p_i > 0$ reflecting the skill rate of each item and constrained the paired comparison probabilities as follows

$$\eta_{ii'} = \frac{p_i}{p_i + p_{i'}}. \tag{2}$$

This is the basic equation of the well-known *Bradley–Terry model* (BTM), which the authors applied only to PC data. Mallows (1957) suggested substituting expression (2) into the BSM, leading to the *Mallows–Bradley–Terry model*. Mallows proposed other simplifications reducing the BSM to specific DBMs, extensively described in Section 2.3. We conclude this subsection by stressing that the BTM corresponds to the TM with only two items whose utilities follow the Gumbel distribution. Fundamental works on models for PCs are Bradley and Terry (1952) and Bradley (1976, 1984), whereas for a detailed review the reader can refer to Cattelan (2012).

2.3. Distance-based models

Roughly speaking, the DBM can be interpreted as the analogue of the normal distribution on the finite discrete space S_K. In fact, it is an exponential location–scale model indexed by a discrete location parameter $\sigma \in S_K$, called *modal* or *consensus ranking*, and a *concentration parameter* $\lambda \in \mathbb{R}_0^+$. Each distribution in the DBM class takes the form

$$\boldsymbol{P}(\pi|\sigma,\ \lambda) = \frac{1}{Z(\lambda)} e^{-\lambda d(\pi,\sigma)} \qquad \pi \in S_K, \tag{3}$$

where $Z(\lambda) = \sum_{\pi \in S_K} e^{-\lambda d(\pi,\sigma)}$ is the normalization constant and $d(\cdot,\ \cdot)$is a metric on S_K. More generally, due to the pioneering work of Mallows (1957), one usually refers to the probability function (3) as the *Mallows model*. The probability mass function (3) is unimodal at $\pi = \sigma$ and decreases symmetrically as the distance from σ increases with a rate calibrated by the concentration λ.

By changing the distance measure d in (3), one can define different families of parametric distributions for ranked data. Examples of the most common metrics for rankings are:

- *Kendall's distance* $d_K(\pi,\sigma)$, equal to the minimum number of adjacent transpositions needed to transform π^{-1} into σ^{-1}.
- *Cayley's distance* $d_C(\pi,\sigma)$, equal to the minimum number of arbitrary transpositions needed to transform π^{-1} into σ^{-1}.
- *Hamming's distance* $d_H(\pi,\sigma)$, equal to the number of items ranked differently in π and σ.
- *Spearman's distance* $d_S(\pi, \sigma) = \sum_{i=1}^{K} (\pi(i) - \sigma(i))^2$.

The determination of $Z(\lambda)$ could be computationally demanding, as it requires the summation over all possible rankings. Nevertheless, some distances such as d_K, d_C and d_H lead to a convenient closed-form expression for $Z(\lambda)$ depending only on λ and K (Fligner & Verducci, 1986).

Popular instances of the DBM class are obtained by adopting the distances d_K or d_S leading, respectively, to the *Mallows* ϕ- or θ-*model*. Actually, Mallows (1957) derived such restricted DBMs in the attempt to simplify the BSM, by setting the special form $\eta_{ii'} = (1 + \tanh((i' - i)\log\theta + \log\phi))/2$ and fixing either $\theta = 1$ or $\phi = 1$. This implies that both Mallows models are nested in the BSM.

2.4. Stagewise models

Stagewise models rely on the assumption that the ranking process can be decomposed into a sequence of $K-1$ independent stages. A fundamental contribution to this class can be found in Fligner and Verducci (1988), introducing a very general family of probability distributions called *multistage models*. It is based on the existence of a true reference ranking in the population, as also assumed in the DBM. Any ranking π can be equivalently expressed in terms of the vector $\boldsymbol{V}(\pi|\sigma) = (V_1(\pi|\sigma),\ \dots,\ V_{K-1}(\pi|\sigma))$ collecting the number of mistakes made by judge π over the $K-1$ stages with respect to the presumed correct ranking σ. By assuming that the V_ts are independent, the model set-up

$$\boldsymbol{P}(\pi|\sigma) = \boldsymbol{P}(\boldsymbol{V}(\pi|\sigma)) = \prod_{t=1}^{K-1} \boldsymbol{P}(V_t(\pi|\sigma) = v_t) \qquad \pi \in S_K \tag{4}$$

is known as the *free model* (FM). Equation (4) represents the most general multistage ranking model, indexed by the choice probabilities $\{\boldsymbol{P}(V_t = v_t) : v_t = 0, \ \ldots, \ K - t \text{ and } t = 1, \ \ldots, \ K - 1\}$. An important subclass of the FM can be obtained by setting an exponential form for the choice probabilities

$$\boldsymbol{P}(V_t = v_t|\lambda_t) = \frac{e^{-\lambda_t v_t}}{\sum_{v=0}^{K-t} e^{-\lambda_t v}} \qquad t = 1, \ \ldots, \ K - 1,$$

leading to the so-called ϕ-*component model*. In a previous work (Fligner & Verducci, 1986), the same authors had already derived the ϕ-component model by starting from a different motivation, specifically a multiparameter extension of the DBM. The starting point is the property of some metrics for rankings, such as d_K and d_C, being decomposable into the sum of $K-1$ independent components,

$$d(\pi, \ \sigma) = \sum_{t=1}^{K-1} V_t(\pi|\sigma), \tag{5}$$

which can be regarded as the factorization of the global distance into the discrepancies over the ranking stages. By applying a non-negative constant λ_t to each term, Fligner and Verducci (1986) further proposed plugging (5) into (3), in order to transfer the stagewise construction to the DBM and derive the *generalized Mallows model* (GMM), given by

$$\boldsymbol{P}(\pi|\sigma, \underline{\lambda}) = \frac{e^{-\sum_{t=1}^{K-1} \lambda_t V_t(\pi|\sigma)}}{Z(\underline{\lambda})} \qquad \pi \in S_K. \tag{6}$$

By recognizing in (6) the product of independent exponential models on the V_ts, the coincidence of the GMM with the ϕ-component model becomes apparent. The equality constraint $\lambda_t = \lambda$ for all $t = 1,\ldots,K-1$ leads directly to the standard DBM with either $d = d_K$ or $d = d_C$.

2.4.1. The Extended Plackett–Luce model

The idea of the ranking process divided into independent stages is shared also by the EPL, an extension of the PL suggested by Mollica and Tardella (2014) and based on the relaxation of the canonical forward order assumption. The EPL is a stagewise model postulating that

$$\boldsymbol{P}_{EPL}(\pi^{-1}|\rho, \ \underline{p}) = \boldsymbol{P}_{PL}(\pi^{-1} \circ \rho|\underline{p}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^{K} p_{\pi^{-1}(\rho(v))}} \qquad \pi^{-1} \in S_K, \tag{7}$$

where the symbol $\circ$ denotes composition between two permutations, the reference order $\rho = (\rho(1), \ \ldots, \ \rho(K)) \in S_K$ is the discrete model parameter and the positive quantities p_is are referred to as *support parameters*. The latter are proportional to the probabilities

$\mathbf{P}(\pi^{-1}(\rho(1)) = i)$ for each item $i = 1,\ldots,K$ to be ranked in the position $\rho(1)$ indicated by the first entry of the reference order ρ. Henceforth, we will refer to model (7) as $EPL(\rho, \underline{p})$ for short. For example, if $K = 4$ and $\rho = (4,1,3,2)$, it means that the ranker follows an alternating attribution of the positions, starting with the specification of the least liked item at the first stage ($\rho(1) = 4$), then the most liked one at the second step ($\rho(2) = 1$), followed by the attribution of the third ($\rho(3) = 3$) and the second position ($\rho(4) = 2$) in the last two stages. This implies that $\mathbf{P}_{\mathrm{EPL}}(\pi^{-1} = (2,3,1,4)) = \mathbf{P}_{\mathrm{PL}}(\pi^{-1} = (4,2,1,3))$.

The popular PL is the special instance of the EPL with the *forward reference order* $\rho_F = (1,2,\ldots,K)$, where the ranker assigns sequentially the positions from the top to the bottom and the order of the ordering entries coincides with that of the item selections. By considering the reversed rank assignment process $\rho_B = (K,K-1,\ldots,1)$, one has another special case of the EPL referred to as *backward PL*.

The crucial difference between the EPL (and hence the PL) and the FM is the following: in the former, the stepwise probabilities are indexed by the available alternatives, whereas in the latter they depend on the stages and the amount of disagreement with respect to the correct choice σ.

Finally, without loss of generality, let us consider the case of $K = 4$ items and the $EPL(\rho, \underline{p} = (.4, .3, .2, .1))$. One can easily verify that the equality

$$\boldsymbol{P}(\pi(1) < \pi(3) | \pi(2) < \pi(4)) = \boldsymbol{P}(\pi(1) < \pi(3) | \pi(4) < \pi(2)),$$

implied by the peculiar property of the TM recalled at the end of Section 2.1, is met only when $\rho = \rho_F$ and $\rho = \rho_B$. This means that, in general, the EPL does not belong to the broad OSM family.

2.4.2. Inference for the EPL

Inference on the EPL and its generalization into a finite mixture framework was originally addressed from the maximum likelihood estimation (MLE) perspective in Mollica and Tardella (2014) via the hybrid expectation-maximization-minorization (EMM) algorithm. Recently, Mollica and Tardella (2021) introduced the Bayesian inference of the EPL, where a discrete uniform distribution for the reference order and independent conjugate gamma densities for the support parameters were chosen for the prior specification. A tuned joint Metropolis-within-Gibbs sampling (TJM-within-GS) was developed to conduct approximate posterior inference on the mixed-type parameter space. The TJM-within-GS was also adapted for the inference of the Bayesian EPL mixtures (Mollica & Tardella, 2019) and for the EPL with order constraints on the reference order (Mollica & Tardella, 2018). The Markov chain Monte Carlo (MCMC) procedure reduces to a GS scheme for the inference on the mixtures of PL (Mollica & Tardella, 2017). Bayesian estimation for the EPL has recently also been considered in Johnson, Henderson, and Boys (2021).

3. Goodness-of-fit diagnostics for ranking models: a review

In the existing reviews of the ranking data literature (Alvo & Yu, 2014; Liu et al., 2019; Marden, 1995), a very limited emphasis is placed on the model assessment issue, testifying to the lack of a systematic and updated overview on the topic and, at the same time, the wide margin of new research directions. In this section we provide a unified and comprehensive review of the model fit diagnostics developed for the analysis of ranking data in both the frequentist and the Bayesian inferential framework, although we will then concentrate on our own original methods from the former estimation perspective.

3.1. Frequentist literature

In the frequentist framework, the assessment of model adequacy for the observed data can be addressed with standard methods such as the likelihood ratio or the chi-squared test for finite discrete distributions. One of the first contributions on evaluating the fit of ranking models was made by Cohen and Mallows (1983), who handled separately the two cases $K < 5$ and $K \geq 5$. In the former situation, the cardinality of the ranking space support is manageable and the $K!$ ranked sequences can be regarded as the categories of a multinomial distribution. Thus, when K is small and N is large enough, model fit can be assessed by appropriately quantifying the dissimilarity between the observed frequencies of each ranking and the expected ones under the estimated model.

However, the rapidly increasing cardinality of the ranking space makes this approach impracticable for larger values of K, due to the possible occurrence of sparse data. In fact, null or low frequencies encountered for some ranking patterns imply that the asymptotic properties of the aforementioned test statistics do not work well in practice. So, for the case $K \geq 5$, a more parsimonious representation of the data is typically needed and Cohen and Mallows (1983) suggested to identify relevant partitions of the permutation set capturing meaningful features of the preference elicitation. In so doing, model fitness can then be evaluated on each subset as previously described for smaller values of K. Some examples are the groupings of the rankings according to their Kendall distance from the estimated modal sequence, formerly proposed by Feigin and Cohen (1978) as a natural diagnostic tool for the ϕ-model. Additionally, Cohen and Mallows (1983) adopted the partition induced by the PCs and employed them to check the adequacy of the Thurstone–Mosteller–Daniels model. For the latter, under the independence assumption, they further computed the standardized deviates to measure the difference between expected and observed frequencies and also displayed the absolute values of the statistics on a half normal probability plot, to better highlight local misfits of the data. A similar method was employed by Yu (2000), who divided the rankings into the subgroups of the sequences with the same item in the top position and compared the sample top frequencies with those expected under the OSM with correlated normal utilities.

Finally, we recall an interesting strand of the goodness-of-fit literature on the TM based on the preliminary transformation of the rankings into the PCs. For example, by exploiting the resulting multivariate binary data structure, Maydeu-Olivares and Böckenholt (2005) revisited the TM within the structural equation modelling framework and proposed to overcome the problem of sparse data by applying the mean adjusted test statistics defined in Satorra and Bentler (1994, 2001). Still in the context of the TM and its possible connections with item response theory, Maydeu-Olivares and colleagues (Maydeu-Olivares, 2001, 2002; Maydeu-Olivares & Brown, 2010; Maydeu-Olivares & Joe, 2006) investigated the properties and usefulness of limited information methods to develop model estimation procedures and diagnostic tools. The sparsity issue is mainly addressed by considering the first- and second-order marginals of the multidimensional contingency tables.

3.2. Bayesian literature

Relevant goodness-of-fit diagnostics appeared in more recent works within the Bayesian ranking literature. In this framework, goodness-of-fit assessment accounts for the randomness of the parameter, rather than relying on a single point estimation as is typical in the frequentist domain. Specifically, the Bayesian approach relies on the construction of a *discrepancy variable*, which depends on both data and parameters and can be employed in the so-called *posterior predictive check*. The core idea is to assess the conformity of the observed value of the discrepancy with its realizations obtained by sampling from the posterior predicted distribution under the estimated model. The computation of the reference distribution of the discrepancy measure under the assumed model is straightforward when a sample from the posterior distribution is available, as in the output of MCMC methods. See Meng (1994) and Gelman, Meng, and Stern (1996) for a general description of Bayesian assessment methods via posterior predictive checks and the more recent works by Hjort, Dahl, and Steinbakk (2006) and Kollenburg, Mulder, and Vermunt (2017) on the calibration of posterior predictive p-values. For model-based ranking analysis, Yao and Böckenholt (1999) focused on paired, triple and quadruple comparisons as empirical summaries. Tsai and Yao (2000) conducted an extensive Monte Carlo simulation study to evaluate the validity of the posterior predictive check for testing the adequacy of alternative OSM. They considered different discrepancy measures and analysed the effect of the number K of alternatives, the sample size N and the type of misspecification on the lack-of-fit detection. In particular, they proved the usefulness of the marginal rank distributions, also known as *first-order marginals*, providing the counts that each item i is ranked in position j. Finally, Mollica and Tardella (2017) constructed two discrepancy variables, based respectively on the top and the PC frequencies, and described how to use them to conduct the posterior predictive check for the Bayesian PL mixture unconditionally and conditionally on the length of the observed partial top rankings.

4. Novel EPL diagnostics

The reviews provided in Section 3 reveal that specific diagnostic tools to evaluate the model adequacy of the class of multistage ranking distributions are very limited and their effectiveness has not been deeply explored and compared. One of the objectives of the present work is to address the goodness-of-fit issue for the EPL specification from the frequentist point of view.

4.1. Inverse monotonicity of the last-stage item probabilities

Let us suppose that $EPL(\rho, \underline{p})$ is the sampling distribution of the ranked observations. Under this model scenario, the marginal item selection probabilities at the first stage are proportional to the support parameters. When the first entry of the reference order is $\rho(1) = 1$, these coincide with the marginal probabilities for each item to be ranked top and, hence, preferred to all the other alternatives. On the other hand, the marginal item selection probabilities at the last stage follow the reverse order of the support parameters. Henceforth, we will refer to this remarkable property as *inverse monotonicity of the last-stage item probabilities*. When K is small, it is rather easy to determine the last-stage item probabilities (see the illustrative example for $K = 3$ in Appendix 1). However, the computational burden needed to exactly compute the marginal item probabilities at each stage represents a non-trivial task that becomes infeasible for larger values of K. The formal proof of the inverse monotonicity of the last-stage item probabilities for any K is, to the best of our knowledge, new and it is provided in Appendix 2. The proof relies on an appropriate method to index the sequences contributing to the construction of the marginal item probabilities, which facilitates the ordinal comparison between them.

4.2. Testing the inverse monotonicity of the last-stage item probabilities

So, if we have some data simulated from $EPL(\rho, \underline{p})$, we expect the marginal frequencies of the items at the first stage to be ranked according to the order of the corresponding support parameter components. On the other hand, we expect the marginal frequencies of the items at the last stage to be ranked according to the reverse order of the corresponding support parameter components. One can then derive that the ranking of the marginal frequencies of the items corresponding to the first and last stage should sum to K+1, no matter what their support is. Of course, this is less likely to happen when the sample size is small or when the support parameter components are not so different. In any case, one can define a test statistic by considering, for each couple of integers (j,j') that may represent the first- and the last-stage ranks, namely $\rho(1)$ and $\rho(K)$, a discrepancy measure $T_{jj'}(\underline{\pi})$ between K+1 (the sum of the expected ranks) and the sum of the observed ranks of the frequencies corresponding to the same item extracted in the first and in the last stage. Formally, let $\underline{r}_j^{[1]} = \left(r_{j1}^{[1]}, \ \ldots, \ r_{jK}^{[1]}\right)$ and $\underline{r}_{j'}^{[K]} = \left(r_{j'1}^{[K]}, \ \ldots, \ r_{j'K}^{[K]}\right)$ be the marginal item frequency distributions for the jth and j'th positions to be assigned, respectively, at the first [1] and last [K] stage. In other words, the generic entry $r_{ji}^{[s]}$ is the number of times that item i is ranked jth at the sth stage. The proposed EPL diagnostic relies on the discrepancy

$$T_{jj'}(\underline{\pi}) = \sum_{i=1}^{K} \left|\text{rank}\left(r_{ji}^{[1]}\right) + \text{rank}\left(r_{j'i}^{[K]}\right) - (K+1)\right|, \tag{8}$$

implying that the smaller the value of $T_{jj'}(\underline{\pi})$, the greater the plausibility that the two integers (j,j') represent the first and the last components of the reference order. In this sense, $T_{jj'}(\underline{\pi})$ can be also reinterpreted as a measure of the closeness of the positions j and j' in the rank attribution path. To globally assess the conformity of the sample with the EPL, we consider the statistic

$$T_m(\underline{\pi}) = \min_{j<j'} T_{jj'}(\underline{\pi}). \tag{9}$$

4.3. Testing the independence of irrelevant alternatives

With the aim of further enlarging the collection of diagnostics of fit for the EPL class, we focus our attention also on a well-known property of the PL. In particular, we consider the distinguishing assumption of the PL known as *Luce's choice axiom* or the *independence of irrelevant alternatives* (IIA) to construct another specification test. The IIA states that the relative preference between two items i and i' does not depend on the liking for the other alternatives belonging to the choice set. For the EPL, the IIA hypothesis implies that the probability ratio of selecting item i over item i' is constant over the stages of the ranking process (constant ratio rule), as long as the two items are both still available. Formally, let $I_{st} = I \backslash \{\pi_s^{-1}(\rho(1)), \ \ldots, \ \pi_s^{-1}(\rho(t-1))\}$ be the choice set composed of the alternatives available at the tth stage for subject s, that is, those items which have not been selected by the ranker s before stage t, and hence removed from the comparison. By introducing the binary indicator

$$\xi_{ii'st} = \begin{cases} 1 & i,i' \in I_{st}, \\ 0 & \text{otherwise}, \end{cases}$$

one can compute the observed PCs at stage t, where item i is selected before item i', as

$$\tau_{ii't} = \sum_{s=1}^{N} \xi_{ii'st} \mathbb{I}_{[\rho^{-1}(\pi_s(i)) < \rho^{-1}(\pi_s(i'))]}.$$

The IIA implies that the expected PC frequency at stage t of choosing item i over item i' is

$$\tau^*_{ii't} = N_{ii't} \frac{p_i}{p_i + p_{i'}},$$

equal to the product of the total number N$_{ii't}$ of PCs between i and i' at stage t, given by

$$N_{ii't} = \tau_{ii't} + \tau_{i'it} = \sum_{s=1}^{N} \xi_{ii'st},$$

and the theoretical PC probability under the EPL, concerning the choice of the two items from the entire set I of the K alternatives. Hence, a chi-squared statistic for the IIA assumption can be defined as

$$X^2_{IIA} = \sum_{t=1}^{K-1} \sum_{i<i'} \frac{(\tau_{ii't} - \tau^*_{ii't})^2}{\tau^*_{ii't}}.$$

The IIA diagnostic operates in a stagewise manner by assessing the relative selection probability of each pair of items (i,i') at each stage $t = 1,\ldots,K-1$ of the ranking process.

5. Comparative assessment of goodness-of-fit diagnostics for ranking models

5.1. Plan of the simulation study

After introducing the novel test statistics (9) and (10), one should inquire into their inferential effectiveness. In order to do so, we first provided an approach for controlling their Type I error rate, and then we investigated their comparative power properties with respect to those of some standard goodness-of-fit tools for ranking models. In the absence of analytical results for the reference distribution under the null hypothesis (EPL assumption), a bootstrap approach was adopted. Specifically, the probability that the test statistic is greater than or equal to the observed value under the EPL assumption was approximated with bootstrap p-values based on 1,000 data sets drawn from the inferred EPL. Deviations from the EPL model should yield greater values of the test statistics than those expected under the model generated, and hence smaller p-values. Finally, for each model adequacy criterion, we estimated the Type I error and correct rejection rates (power) by the proportion of the times that the p-value was smaller than or equal to the nominal .05 critical threshold, depending on the true generating model.

A simulation study was conducted under alternative model specifications, involving the comparison with the chi-squared statistics based on the marginal top selection frequencies, the first-order marginals and the PCs, given respectively by

$$X^2_{TOP} = \sum_{i=1}^{K} \frac{(m_{1i} - m^*_{1i})^2}{m^*_{1i}} \qquad X^2_M = \sum_{t=1}^{K}\sum_{i=1}^{K} \frac{(m_{ti} - m^*_{ti})^2}{m^*_{ti}} \qquad X^2_{PC} = \sum_{i<i'} \frac{(\tau_{ii'} - \tau^*_{ii'})^2}{\tau^*_{ii'}}.$$

The marginal expected frequencies were obtained as $m^*_{1i} = Np_i$, $\tau^*_{ii'} = Np_i/(p_i + p_{i'})$ whereas, due to complexity of their exact computation, the m^*_{ti} were estimated by Monte Carlo simulation.

Finally, note that X^2_M is a stagewise extension of the classical chi-squared statistic X^2_{TOP}. In fact, the latter is obtained from X^2_M by considering only the term $t = 1$ of the outer sum, concerning the marginal item distribution in the top stage, that is,

$$m_{1i} = \sum_{s=1}^{N} \mathbb{I}_{[\pi_s^{-1}(\rho(1))=i]}.$$

Similarly, the IIA diagnostic can be regarded as a stagewise generalization of X^2_{PC}. For the latter, the comparison between item i and i' is considered only at the first stage, that is, in the context of the whole item set I for which $N_{ii'1} = N$. Finally, note that X^2_{PC} is based on a quadratic form of the PC frequencies and coincides with the first-order moment statistic introduced in Maydeu-Olivares and Böckenholt (2005) in the general context of limited information diagnostics for multivariate binary data. It is worth mentioning that the adaptation of the limited information diagnostics introduced in Maydeu-Olivares and Böckenholt (2005) to the context of ranking data analysis is not straightforward. In any case, although it could lead to better control of the asymptotic properties of the test statistics, we cannot pursue this direction since, for the mixed-type (continuous and discrete) EPL parametrization, we cannot rely on the required best asymptotic normal property of the MLE.

5.2. Simulation study

A comparative evaluation of the diagnostic tools was carried out by means of an extensive simulation study. For each possible combination (K,N), with values varying respectively in the grids $K \in \{5,\ 10,\ 20,\ 40\}$ and $N \in \{300,\ 450,\ 600\}$, we drew 100 data sets with N orderings of K items from the following ranking distributions:

- EPL;
- DBM with the Kendall distance (DBM-Kend);
- DBM with the Cayley distance (DBM-Cay);
- DBM with the Hamming distance (DBM-Ham);
- TH with independent normal latent scores (TH-norm), corresponding to the Case III model.

The true parameter values of the above model scenarios were generated according to the following schemes: (i) $\rho \sim Unif\{S_K\}$ and $p_i \overset{iid}{\sim} Unif(0,1)$ for the EPL; (ii) $\sigma_0 \sim Unif\{S_K\}$ and $\lambda \sim Unif(0,3)$ for the modal ranking and the concentration parameter of the three DBMs considered; and (iii) $\mu_i, \quad \sigma_i \overset{iid}{\sim} Unif(0,1)$ for the means and the standard deviations of the latent item scores of the TH-norm.

The estimated Type I error rates and power are reported in Appendix 4 and can be more easily evaluated in Figures 1 and 2. The simulation study revealed satisfactory performance of all the diagnostics considered regarding Type I error rates as long as $K \leq 20$. In fact, in all these settings, they were below .05. However, for $K = 40$, some troublesome deviations from the nominal level were detected for $X^2_{TOP}, X^2_M, X^2_{PC}$ and X^2_{IIA}. For X^2_{IIA}, the departure from .05 is less remarkable, whereas the T_m diagnostic is consistently under the .05 threshold for all sample sizes considered.

On the other hand, noteworthy differences among the statistics emerged in terms of power. Firstly, T_m exhibited consistently lower performance of the estimated power under almost all the model scenarios considered. However, this diagnostic is the only one which can safely be used with the largest value $K = 40$, since it takes the Type I error rate under control. Hence, its estimated power is not overestimated, as in the case of the other diagnostics which exceed the nominal rejection rate under the null hypothesis (EPL assumption). At least two motivations can be put forward to argue the evidence of lower power of T_m. The first is related to its formal definition; in fact, this is a parameter-free measure based on the ranks of the expected marginal frequencies, rather than on the computation of the parameter-dependent first- and last-stage theoretical probabilities. This makes T_m by construction a rougher diagnostic in the comparison with the other statistics. Secondly, its remarkably low power for the DBM-Kend suggested that the monotonicity property of the first- and last-stage item probabilities is not specific to the EPL, but it is shared by other rankings models. This implies that the T_m statistic could not

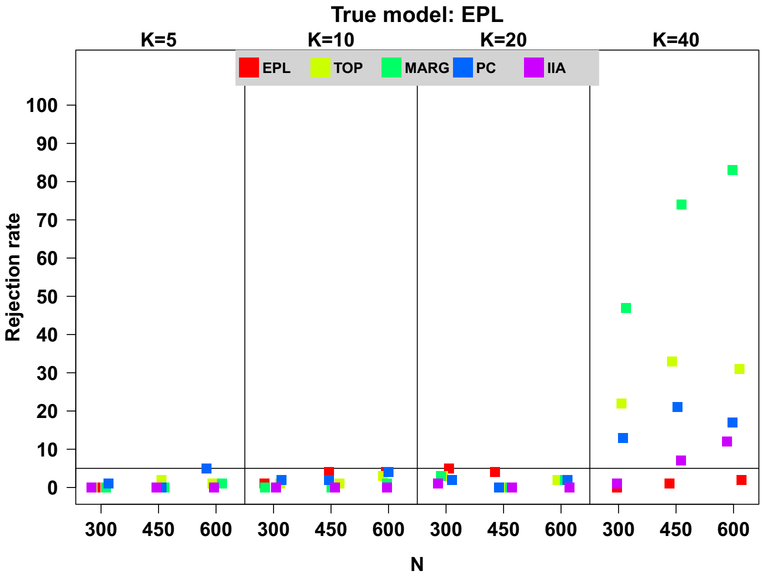

Figure 1. Type I error rates (%) of alternative goodness-of-fit diagnostics under the EPL assumption. The diagnostic tools considered are $T_m, X^2_{TOP}, X^2_M, X^2_{PC}$ and X^2_{IIA}, respectively labelled in the legend as EPL, TOP, MARG, PC and IIA. The reference distributions of all the test statistics under the EPL assumption have been approximated with the bootstrap method.

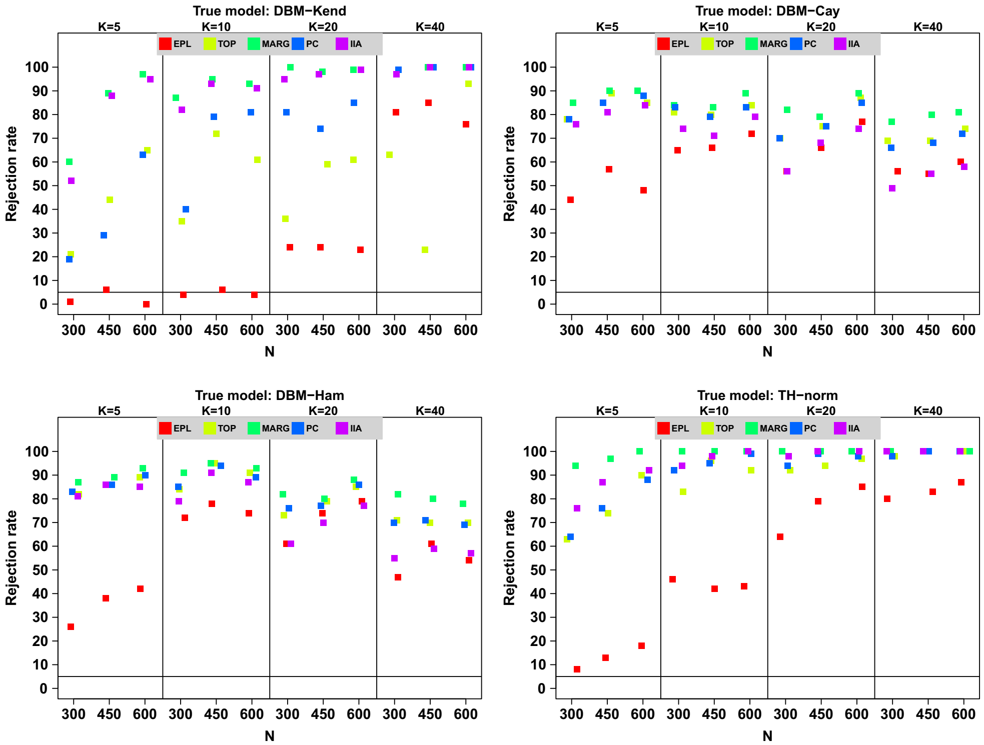

Figure 2. Rejection rates (%) of the EPL assumption for alternative goodness-of-fit diagnostics computed on simulated data from different model scenarios. The diagnostic tools considered are T_m, X^2_{TOP}, X^2_M, X^2_{PC} and X^2_{IIA}, respectively labelled in the legend as EPL, TOP, MARG, PC and IIA. The reference distributions of all the test statistics under the EPL assumption have been approximated with the bootstrap method.

discriminate the EPL from other parametric families with sufficient flexibility to describe an underlying stagewise elicitation process with a certain coherence, over the stages, about the preferences of the items. This is also the case for some subclasses of the TM. In fact, besides the trivial case of the OSM with Gumbel distributions for latent utilities (corresponding to the PL), the property is recovered also when adopting independent normals with varying means and constant variances (Case V model). Conversely, the property does not hold in general for the DBMs with any metric other than the Kendall. Although an exact computation of the marginal item distributions at each stage is a difficult task (see, for example, the case of the $EPL(\rho, \underset{\sim}{p})$ in Appendix 2), these claims can easily be verified for a specific ranking model via a simulation approach. In fact, here we recorded evidence of better power performance for T_m with the use of the Cayley and Hamming distances.

Another consistent piece of evidence highlighted by the comparative analysis concerns the best-performing diagnostic, which turned out to be the one relying on the first-order marginals. However, it is no less apparent that, for higher values of *K* and *N*, the performance of the new IIA statistic under DBM-Ken and TH-norm is pretty much equivalent to that of X^2_M and, in general, always better than the remaining competing statistics typically used in real-data applications. For DBM-Cay and DBM-Ham, on the other hand, its performance is close to that of the T_m statistic. So, by noting also that the X^2_{IIA} seems to suffer less from lack of control of the Type I error rate for $K = 40$ than the generic ranking model test statistics, from the simulation study an overall positive verdict can be given about the novel IIA diagnostic.

6. Illustrative applications of the goodness-of-fit diagnostics

6.1. Application to the salad data set

As a first application, we considered the salad data set available in the *prefmod* package (Hatzinger & Dittrich, 2012) in R, containing $N = 32$ rankings of $K = 4$ types of salad dressings compared in terms of the perceived level of tartness.

As displayed in Table 1, all the test statistics considered turned out to be well above the critical .05 threshold, and hence did not highlight any issue of lack-of-fit for the estimated EPL in terms of the descriptive summaries considered.

6.2. Application to the occupation data set

For the second application, we focused on a data set with higher values of both K and N, specifically, the occupation data set available in the *PLMIX* package (Mollica & Tardella, 2020) in R. This came from a survey conducted on graduates from the Technion-Israel Institute of Technology. A sample of $N = 143$ graduates were asked to rank $K = 10$ professions according to their perceived prestige: 1 = faculty member, 2 = owner of a business, 3 = applied scientist, 4 = operations researcher, 5 = industrial engineer, 6 = manager, 7 = mechanical engineer, 8 = supervisor, 9 = technician, 10 = foreman.

With the sole exception of the X^2_{TOP} diagnostic based on the frequencies of selection at the first stage, all the available diagnostics suggested the rejection of the EPL assumption (Table 1). We then further explored the usefulness of the EPL goodness-of-fit diagnostics to gain further insights into the same data set. For illustrative purposes we inspected the possible presence of a group structure in the sample by estimating finite EPL mixtures with a varying number of components ($G = 1,2,3,4$) through the iterative EMM procedure described in Mollica and Tardella (2014) and compared them using the Bayesian information criterion (BIC) introduced by Schwarz (1978). A mixture of EPLs with two groups was selected as the optimal model (BIC = 2,959.37) with, respectively, weights .72 and .28, support parameters (0.002,0.001,0.003,0.009,0.011,0.010,0.026,0.018,0.390, 0.531) and (0.466,0.039,0.270,0.095,0.077,0.009,0.039,0.004,0.001,0.000) and almost opposite reference orders (10,9,8,7,5,6,4,3,2,1) and (1,2,3,5,4,6,7,8,9,10). We then assessed the goodness of fit of the EPL on both subgroups of data separately. As displayed in Table 1, there is less critical evidence of EPL misfit when we separately consider the two subgroups for which two distinct reference orders are estimated.

Table 1. *p*-values of the goodness-of-fit diagnostics for the salad and occupation data sets

	Salad	Occupation (full sample)	Occupation (group 1)	Occupation (group 2)
T_m	1.000	0.034	0.126	0.288
X^2_{TOP}	0.631	0.164	0.201	0.509
X^2_{M}	0.739	0.007	0.042	0.138
X^2_{PC}	0.775	0.000	0.021	0.187
X^2_{IIA}	0.880	0.025	0.416	0.603

7. Likelihood-free estimation of the reference order

In this section we explore the utility of the statistic T_m from the inferential point of view. We show how one can quickly obtain from the observed rankings a candidate best-fitting reference order relying on the observation that $T_{jj'}(\underline{\pi})$ can be considered as a measure of proximity between the pair of positions j and j' in the unknown sequence ρ.

7.1. The novel heuristic method

Let $\boldsymbol{T}(\underline{\pi}) = (T_{jj'}(\underline{\pi}))$ be the $K \times K$ matrix with entries defined in (8). The computation of $\boldsymbol{T}(\underline{\pi})$ is illustrated by an example reported in Appendix 3. For each component $T_{jj'}(\underline{\pi})$, we have

$$T_{jj'}(\underline{\pi}) \leq u_K,$$

where the upper bound u_K corresponds to the constant value in the main diagonal, that is,

$$u_K = T_{jj}(\underline{\pi}) = \sum_{l=1}^{K} |2l-(K+1)| = 2\left(\sum_{l=1}^{(K-1)/2} 2l\right)^{K \bmod 2}\left(\sum_{l=0}^{K/2-1} (2l+1)\right)^{1-K \bmod 2}.$$

This means that the maximum value in $\boldsymbol{T}(\underline{\pi})$ depends on the observed data only through K: for K odd, u_K is the double sum of the first (K+1)/2 even numbers (starting from zero); for K even, u_K is the double sum of the first K/2 even numbers.

Our heuristic method to estimate the unknown parameter ρ consists of the following steps:

1. Compute

$$\boldsymbol{D}(\underline{\pi}) = |\boldsymbol{T}(\underline{\pi}) - u_K \boldsymbol{J}_K|,$$

where $\boldsymbol{J}_K$ is a $K \times K$ matrix consisting entirely of 1s, so that each component $D_{jj'}(\underline{\pi})$ can be interpreted as a measure of the distance between positions j and j' in the sequential rank assignment process.

2. Use the matrix $\boldsymbol{D}(\underline{\pi})$ as the input of a principal component analysis (PCA).

3. Estimate ρ by taking the non-decreasing ordering of the scores $(\zeta_1, \ldots, \zeta_K)$ of the K positions on the first principal component, given by

$$\hat{\rho} = (\hat{\rho}(1), \ldots, \hat{\rho}(K)) : \zeta_{\hat{\rho}(1)} \leq \ldots \leq \zeta_{\hat{\rho}(K)}.$$

7.2. Effectiveness of the heuristic method

The inferential effectiveness of the proposal to recover the true discrete parameter was explored by means of a simulation study with a varying cardinality K of the item set and sample size N. For each possible combination (K,N), where $K \in \{5, 10, 15, 20, 40\}$ and $N \in \{50, 200, 1000, 10000\}$, we drew 100 data sets $\underline{\pi}^{-1}_{(R)}$ with $R = 1,\ldots,100$ from the EPL according to the scheme described in Section 5.2. For comparison purposes, we inferred the reference order of each simulated sample with: (i) the heuristic strategy described in Section 7.1; (ii) the same heuristic scheme with the PCA replaced by multidimensional scaling (MDS); and (iii) the MLE approach via the EMM algorithm proposed in Mollica and Tardella (2014), which is considered as the reference method for the present estimation task. Finally, the estimation performance of the competing strategies was compared in terms of:

- % recoveries $= \sum_{R=1}^{100} \mathbb{I}_{[\rho_{(R)} \in \hat{\rho}^{equiv}_{(R)}]}$, the percentage of times that the actual reference order belongs to the equivalence class corresponding to the estimated reference order. Two distinct reference orders ρ and ρ′ are considered equivalent with respect to the distance matrix $\boldsymbol{D}(\underline{\pi})$ if, for all $t = 1,\ldots,K$, either $\rho(\mathrm{t}) = \rho\prime(\mathrm{t})$ or $D_{\rho(t)\rho'(t)}(\underline{\pi}) = 0$.
- $\bar{r}(\rho^{-1}, \hat{\rho}^{-1}) = \frac{1}{100}\sum_{R=1}^{100} r\left(\rho^{-1}_{(R)}, \hat{\rho}^{-1}_{(R)}\right)$, the average rank correlation coefficient.

The results are shown in Table 2. It is evident that PCA and MDS exhibited essentially the same ability. Compared with the MLE, the heuristic methods exhibited very good results. The percentage of matching consistently grows with *N* and, by also checking the cases where there is not an exact correspondence, on average an analogous trend is found for the correlation. Additionally, if we look at a fixed *N*, the percentage of recoveries shows a worse tendency for larger values of *K*. In this regard, the cases $K \geq 10$, combined with a relatively very low ($N = 50$) and very high ($N = 10{,}000$) sample size, deserve some considerations to stress typical issues which can be encountered in a ranking data analysis. First, in a sparse data situation, all of the estimation techniques exhibit great uncertainty in exactly recovering the actual ρ, as testified by the negligible values of the recovery

Table 2. Inferential performance of the heuristic methods via PCA and MDS in estimating the reference order on simulated data compared to the MLE via EMM algorithm

	% recoveries			$\bar{r}(\rho^{-1}, \hat{\rho}^{-1})$		
(*K*,*N*)	PCA	MDS	MLE	PCA	MDS	MLE
(5,50)	54	50	54	0.57	0.49	0.76
(5,200)	76	73	86	0.65	0.66	0.95
(5,1000)	90	87	98	0.85	0.78	1.00
(5,10000)	97	98	–	0.92	0.94	–
(10,50)	2	3	4	0.56	0.56	0.90
(10,200)	18	16	25	0.83	0.86	0.96
(10,1000)	47	49	74	0.94	0.92	0.99
(10,10000)	77	76	–	0.96	0.96	–
(15,50)	0	0	0	0.83	0.83	0.91
(15,200)	1	1	3	0.79	0.83	0.97
(15,1000)	24	25	44	0.97	0.97	1.00
(15,10000)	68	73	–	0.97	0.97	–
(20,50)	0	0	0	0.81	0.76	0.92
(20,200)	0	0	0	0.91	0.91	0.98
(20,1000)	2	0	11	0.97	0.95	0.99
(20,10000)	22	19	–	0.98	0.98	–
(40,50)	0	0	–	0.86	0.89	–
(40,200)	0	0	–	0.95	0.95	–
(40,1000)	0	0	–	0.97	0.98	–
(40,10000)	0	0	–	0.98	0.98	–

percentage. From a computational perspective, although a better behaviour of the MLE is expected for $N = 10{,}000$, this has not been implemented since, without a specialized program, fitting the EPL to a large sample can be extremely demanding, if not downright infeasible. This is especially true for the more flexible EPL class, due to the impact of the reference order on the normalizing term of the likelihood and the need for its iterative update during the optimization procedure (Mollica & Tardella, 2014). Moreover, the computational burden is further aggravated by the multiple initializations needed to address the issue of local maxima. Of course, even if one increases the number of starting values, the fast-growing dimension of the reference order space S_K can make the multiple starting point strategy rapidly ineffective for larger values of K to reach the global maximum. In the light of these remarks on the MLE, the likelihood-free approach can be motivated as a straightforward method that can be combined with the MLE procedure or with an MCMC method in the Bayesian estimation framework. In fact, at a limited computational cost, it can be implemented as a preliminary step of the inferential process to obtain a promising initialization, which can guide and speed up the parameter space exploration towards the achievement of the global optimum, resulting in substantial time savings.

8. Conclusions

After a careful review of the existing literature on ranking data models diagnostics, in this work we presented new methods for improving the analysis of ranking data under the assumption that the observations were generated from the stagewise EPL distribution. In particular, we focused on the lack of specific goodness-of-fit statistics for multistage ranking models and on the peculiar issue related to the EPL concerning inference on the discrete parameter component.

Inspired by two formal properties of the EPL parametric class, one discussed and proven for the first time in the present work and the other inherited from the PL subclass (the IIA assumption), we constructed and explored the usefulness of two novel sample statistics to test the appropriateness of the EPL distribution. The comparative performance of the two diagnostics with respect to more general goodness-of-fit tests for ranking models was evaluated by means of a simulation study under alternative data-generating models. On the one hand, the comparison highlighted the lower power of the statistics based on the property of inverse monotonicity of the last-stage item probabilities to distinguish the EPL from the other distributions. On the other hand, this turned out to be the only statistic which preserves Type I error control for all the range of number of items considered in the simulation study. The simulation study also identified the generic test statistic based on the first-order marginals as the best-performing one, although for larger values of K and N the proposed IIA diagnostic exhibited equivalent power. We stress that, differently from X^2_M, the novel X^2_{IIA} represents a specific test for the EPL assumption. In our opinion, the higher power of the two statistics could depend on a better account of the K-dimensional ranking process, that is, the ability of the two statistics X^2_{IIA} and X^2_M to span the whole multivariate dependence structure, rather than only univariate or bivariate marginal features of the preference elicitation, such as the tests based on the top selection frequencies or on the PCs. In this sense, the originality of the simulation results under the EPL specification could stimulate future research on the critical issue concerning the evaluation of the adequacy of ranking models.

We then revisited the usefulness of the property of inverse monotonicity of the last-stage item probabilities from the inferential perspective, as the core ingredient of a heuristic method to estimate ρ. The aim was to address the estimation issue with lower computational costs, by returning a promising sample-based evaluation of ρ that can be used as a good initialization of iterative inferential procedures. The utility of the proposal was checked with a comparative simulation study, which highlighted a satisfactory inferential ability to get consistently close to the true underlying reference order, although outperformed by the MLE (as expected). Hence, the new likelihood-free strategy could fruitfully replace or complement the more conventional and time-consuming multiple-initialization procedure to attain the global optimum of the likelihood. We remark that, by comparing the computing times of all the settings considered, the execution of the procedure scales linearly with N, while being cubic in the number of items. Despite the fact it does not scale well with K, the overall computation cost of the heuristic procedure remains acceptable: it can be executed with $K = 100$ and $N = 10{,}000$ in slightly more than 15 minutes, while the MLE would take almost 2 days. In carrying out the goodness-of-fit procedure, the MLE is always the most computationally demanding component of the whole procedure. While the MLE scales linearly with the sample size and is quadratic in the number of items, the implementation of the goodness-of-fit tests scales better overall, hence it does not represent an issue. Despite the fact that the diagnostic based on the inverse monotonicity does not compare favourably with that relying on the IIA assumption in terms of power, the heuristic method derived from it may be of interest *per se* as a descriptive tool for measuring and qualifying the presence of some deviations from the canonical assignment of ordered positions during the sequential elicitation process.

As a possible future development, we would like to continue with the introduction and evaluation of other specific goodness-of-fit tests for the class of stagewise models, in order to gain further improvement over standard ranking model diagnostics. In particular, the extension to the finite mixture framework would be an important enhancement to address model checking for more flexible ranking data models. Goodness-of-fit diagnostics in the presence of partial observations is also a topic which deserves to be further developed. In fact, to our knowledge, only in the Bayesian literature there are a few contributions in this direction (Johnson, Henderson, & Boys, 2020; Mollica & Tardella, 2017). To this end, the statistics considered would initially require the extension of the MLE for the EPL on samples including partial rankings. Another valuable direction of research could be the Bayesian extension of the novel diagnostic tools allowing for model adequacy evaluation via posterior predictive checks.

Acknowledgement

Open Access Funding provided by Universita degli Studi di Roma La Sapienza within the CRUI-CARE Agreement. [Correction added on 23 May 2022, after first online publication: CRUI-CARE funding statement has been added.]

Conflicts of interest

All authors declare no conflict of interest.

Data availability statement

Data available on request from the authors.

References

Alvo, M., & Yu, P. L. (2014). *Statistical methods for ranking data*. New York, NY: Springer.

Babington Smith, B. (1950). Discussion of professor Ross's paper. *Journal of the Royal Statistical Society: Series B*, *12*(1), 53–56.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics*, *32*, 213–232.

Bradley, R. A. (1984). *Paired comparisons: some basic procedures and examples, Nonparametric methods*, 299–326. Amsterdam: North-Holland.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, *27*, 412–433.

Cohen, A., & Mallows, C. L. (1983). Assessing goodness of fit of ranking models to data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*, 361–374.

Critchlow, D. E., Fligner, M. A., & Verducci, J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology*, *35*, 294–318.

Daniels, H. E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *12*, 171–191.

Feigin, P. D., & Cohen, A. (1978). On a model for concordance between judges. *Journal of the Royal Statistical Society: Series B (Methodological)*, *40*, 203–213.

Fligner, M. A., & Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *48*, 359–369.

Fligner, M. A., & Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association*, *83*, 892–901.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–760.

Gormley, I. C., & Murphy, T. B. (2008). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, *2*, 1452–1477.

Gormley, I. C., & Murphy, T. B. (2010). *Clustering ranked preference data using sociodemographic covariates, Choice Modelling: The State-of-the-Art and the State-of-Practice: Proceedings from the Inaugural International Choice Modelling Conference, 543–569*. Emerald, UK.

Hatzinger, R., & Dittrich, R. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, *48*(10), 1–31.

Henderson, D. A., & Kirrane, L. J. (2018). A comparison of truncated and time-weighted plackett-luce models for probabilistic forecasting of formula one results. *Bayesian Analysis*, *13*, 335–358.

Henery, R. J. (1981). Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(1), 86–91.

Henery, R. J. (1983). Permutation probabilities for Gamma random variables. *Journal of Applied Probability*, *20*, 822–834.

Hjort, N. L., Dahl, F. A., & Steinbakk, G. H. (2006). Post-processing posterior predictive p-values. *Journal of the American Statistical Association*, *101*, 1157–1174.

Johnson, S. R., Henderson, D. A., & Boys, R. J. (2020). Revealing subgroup structure in ranked data using a Bayesian WAND. *Journal of the American Statistical Association*, *115*, 1888–1901.

Johnson, S. R., Henderson, D. A., & Boys, R. J. (2021). On Bayesian inference for the extended Plackett-Luce model. *Bayesian Analysis*, *1*(1), 1–26.

Lee, P. H., & Yu, P. L. H. (2012). Mixtures of weighted distance-based models for ranking data with applications in political studies. *Computational Statistics & Data Analysis*, *56*, 2486–2500.

Liu, Q., Crispino, M., Scheel, I., Vitelli, V., & Frigessi, A. (2019). Model-based learning from preference data. *Annual Review of Statistics and Its Application*, *6*, 329–354.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: John Wiley & Sons Inc.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, *44*, 114–130.

Marden, J. I. (1995). *Analyzing and modeling rank data, Vol. 64 of monographs on statistics and applied probability*. New York, NY: Chapman & Hall.

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*, 209–228.

Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data. *Psychometrika*, *43*, 467–483.

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*, 285–304.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *71*, 935–974.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*, 1142–1160.

Mollica, C., & Tardella, L. (2014). Epitope profiling via mixture modeling of ranked data. *Statistics in Medicine*, *33*, 3738–3758.

Mollica, C., & Tardella, L. (2017). Bayesian mixture of Plackett-Luce models for partially ranked data. *Psychometrika*, *82*, 442–458.

Mollica, C., & Tardella, L. (2018). *Constrained Extended Plackett-Luce model for the analysis of preference rankings, Book of Short Papers – SIS2018: 49th Scientific meeting of the Italian Statistical Society*. 480–486. Pearson, Palermo.

Mollica, C., & Tardella, L. (2019). *Modelling unobserved heterogeneity of ranking data with the Bayesian mixture of Extended Plackett-Luce models, Book of Short Papers – CLADAG 2019: 12th Scientific Meeting of the Classification and Data Analysis Group*. 346–349. Cassino, Italy.

Mollica, C., & Tardella, L. (2020). PLMIX: An R package for modelling and clustering partially ranked data. *Journal of Statistical Computation and Simulation*, *90*, 925–959.

Mollica, C., & Tardella, L. (2021). Bayesian analysis of ranking data with the extended Plackett-Luce model. *Statistical Methods and Applications*, *30*(1), 175–194.

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, *16*(1), 3–9.

Overland, I., & Juraev, J. (2019). Algorithm for producing rankings based on expert surveys. *Psychometrika*, *12*(1), 19.

Plackett, R. L. (1968). Random permutations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*, 517–534.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage Publications Inc.*****

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Stern, H. (1990). Models for distributions on permutations. *Journal of the American Statistical Association*, *85*, 558–564.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Reviews*, *34*, 273–286.

Tsai, R. C., & Yao, G. (2000). Testing Thurstonian Case V ranking models using posterior predictive checks. *British Journal of Mathematical and Statistical Psychology*, *53*, 275–292.

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2017). Posterior calibration of posterior predictive p-values. *Psychological Methods*, *22*, 382–396.

Wang, C., & Bier, V. M. (2013). Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, *61*, 372–385.

Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, *52*(1), 79–92.

Yu, P. L. H. (2000). Bayesian analysis of order-statistics models for ranking data. *Psychometrika*, *65*, 281–299.

Received 21 September 2020; revised version received 2 November 2021

Appendix 1: Inverse monotonicity of the last-stage item probabilities with K = 3 items

Let us consider the simple case with $K = 3$ items. Without loss of generality, we assume that the data-generating mechanism is $EPL(\rho,\ \underline{p})$ with $p_1 \leq p_2 \leq p_3$ and ρ=(1,2,3). The marginal probability for each item to be selected at the first stage can be computed as follows

$$\begin{aligned}\boldsymbol{P}_{EPL}\left(\pi^{-1}(1) = 1|\rho,\ \underline{p}\right) &= \boldsymbol{P}_{PL}(\pi^{-1} = (1,2,3)|\underline{p}) + \boldsymbol{P}_{PL}(\pi^{-1} = (1,3,2)|\underline{p})\\ &= \frac{p_1}{p_1+p_2+p_3}\frac{p_2}{p_2+p_3} + \frac{p_1}{p_1+p_2+p_3}\frac{p_3}{p_2+p_3}\\ &= \frac{p_1}{p_1+p_2+p_3} \propto p_1,\end{aligned}$$

$$\begin{aligned}\boldsymbol{P}_{EPL}\left(\pi^{-1}(1) = 2|\rho,\ \underline{p}\right) &= \boldsymbol{P}_{PL}(\pi^{-1} = (2,1,3)|\underline{p}) + \boldsymbol{P}_{PL}(\pi^{-1} = (2,3,1)|\underline{p})\\ &= \frac{p_2}{p_1+p_2+p_3}\frac{p_1}{p_1+p_3} + \frac{p_2}{p_1+p_2+p_3}\frac{p_3}{p_1+p_3}\\ &= \frac{p_2}{p_1+p_2+p_3} \propto p_2,\end{aligned}$$

$$\begin{aligned}\boldsymbol{P}_{EPL}\left(\pi^{-1}(1) = 3|\rho,\ \underline{p}\right) &= \boldsymbol{P}_{PL}(\pi^{-1} = (3,1,2)|\underline{p}) + \boldsymbol{P}_{PL}(\pi^{-1} = (3,2,1)|\underline{p})\\ &= \frac{p_3}{p_1+p_2+p_3}\frac{p_1}{p_1+p_2} + \frac{p_3}{p_1+p_2+p_3}\frac{p_2}{p_1+p_2}\\ &= \frac{p_3}{p_1+p_2+p_3} \propto p_3,\end{aligned}$$

implying

$$\boldsymbol{P}_{EPL}\left(\pi^{-1}(1) = 1|\rho,\ \underline{p}\right) \leq \boldsymbol{P}_{EPL}(\pi^{-1}(1) = 2|\rho,\ \underline{p}) \leq \boldsymbol{P}_{EPL}(\pi^{-1}(1) = 3|\rho,\ \underline{p}).$$

The marginal probability for each item to be selected at the third (last) stage can be computed as follows

$$\begin{aligned}
\boldsymbol{P}_{EPL}\left(\pi^{-1}(3)=1|\rho,\ \underline{p}\right) &= \boldsymbol{P}_{PL}(\pi^{-1}=(2,3,1)|\underline{p})+\boldsymbol{P}_{PL}(\pi^{-1}=(3,2,1)|\underline{p})\\
&=\frac{p_2}{p_1+p_2+p_3}\frac{p_3}{p_1+p_3}+\frac{p_3}{p_1+p_2+p_3}\frac{p_2}{p_1+p_2}\\
&=\frac{p_2p_3p_1}{p_1+p_2+p_3}\frac{1}{p_1}\left(\frac{1}{p_1+p_3}+\frac{1}{p_1+p_2}\right),
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{P}_{EPL}\left(\pi^{-1}(3)=2|\rho,\ \underline{p}\right) &= \boldsymbol{P}_{PL}(\pi^{-1}=(1,3,2)|\underline{p})+\boldsymbol{P}_{PL}(\pi^{-1}=(3,1,2)|\underline{p})\\
&=\frac{p_1}{p_1+p_2+p_3}\frac{p_3}{p_2+p_3}+\frac{p_3}{p_1+p_2+p_3}\frac{p_1}{p_1+p_2}\\
&=\frac{p_1p_3p_2}{p_1+p_2+p_3}\frac{1}{p_2}\left(\frac{1}{p_2+p_3}+\frac{1}{p_1+p_2}\right),
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{P}_{EPL}\left(\pi^{-1}(3)=3|\rho,\ \underline{p}\right) &= \boldsymbol{P}_{PL}(\pi^{-1}=(1,2,3)|\underline{p})+\boldsymbol{P}_{PL}(\pi^{-1}=(2,1,3)|\underline{p})\\
&=\frac{p_1}{p_1+p_2+p_3}\frac{p_2}{p_2+p_3}+\frac{p_2}{p_1+p_2+p_3}\frac{p_1}{p_1+p_3}\\
&=\frac{p_1p_2p_3}{p_1+p_2+p_3}\frac{1}{p_3}\left(\frac{1}{p_2+p_3}+\frac{1}{p_1+p_3}\right).
\end{aligned}$$

Since $p_1 \leq p_2$, we have

$$\frac{1}{p_1}\left(\frac{1}{p_1+p_3}+\frac{1}{p_1+p_2}\right) \geq \frac{1}{p_2}\left(\frac{1}{p_2+p_3}+\frac{1}{p_1+p_2}\right),$$

implying

$$\boldsymbol{P}_{EPL}(\pi^{-1}(3)=1|\rho,\ \underline{p}) \geq \boldsymbol{P}_{EPL}(\pi^{-1}(3)=2|\rho,\ \underline{p}).$$

Since $p_2 \leq p_3$, we have

$$\frac{1}{p_2}\left(\frac{1}{p_2+p_3}+\frac{1}{p_1+p_2}\right) \geq \frac{1}{p_3}\left(\frac{1}{p_2+p_3}+\frac{1}{p_1+p_3}\right),$$

implying

$$\boldsymbol{P}_{EPL}(\pi^{-1}(3)=2|\rho,\underline{p}) \geq \boldsymbol{P}_{EPL}(\pi^{-1}(3)=3|\rho,\underline{p}),$$

and hence

$$\boldsymbol{P}_{EPL}(\pi^{-1}(3)=1|\rho,\underline{p}) \geq \boldsymbol{P}_{EPL}(\pi^{-1}(3)=2|\rho,\underline{p}) \geq \boldsymbol{P}_{EPL}(\pi^{-1}(3)=3|\rho,\underline{p}).$$

Appendix 2: Proof of the inverse monotonicity of the last-stage item probabilities

Without loss of generality, let $EPL(\rho, \underline{p})$ with $p_1 \leq \ldots \leq p_K$ be the data-generating mechanism. We denote by $q_l^{[t]}$ the probability that item l is selected at stage t under the above $EPL(\rho, \underline{p})$. Hence, the vector $\left(q_1^{[t]}, \ \ldots, \ q_l^{[t]}, \ \ldots, \ q_K^{[t]}\right)$ is the marginal item distribution at stage t. Our interest is in determining the ordering of the probability masses $\left(q_1^{[K]}, \ \ldots, \ q_l^{[K]}, \ \ldots, \ q_K^{[K]}\right)$ relative to the marginal item distribution at the last stage K. To simplify the notation in the proof, we adopt the following conventions: for $t = 1,\ldots,K$, we denote by $i_t = \pi^{-1}(\rho(t))$ the label of the item selected at stage t, implying $q_l^{[t]} = \boldsymbol{P}_{EPL}(i_t = l|\rho, \underline{p})$, and we denote by $\underline{p}_{[D]} = \sum_{i \in D} p_i$ the restricted sum of the support parameters of the items belonging to the choice set $D \subseteq I$. Let us write the marginal probability for item 1 to be chosen in the final step K of the ranking process. This can be obtained by marginalizing out the entries of the previous $K-1$ stages, that is,

$$\begin{aligned}
q_1^{[K]} &= \mathbf{P}_{\mathrm{EPL}}(i_K = 1|\rho, \underline{p}) = \mathbf{P}_{\mathrm{PL}}(i_K = 1|\underline{p}) \\
&= \sum_{i_1 \in I\setminus\{1\}} \cdots \sum_{i_t \in I\setminus\{1, i_1, \ldots, i_{t-1}\}} \cdots \sum_{i_{K-1} \in I\setminus\{1, i_1, \ldots, i_{K-2}\}} \mathbf{P}_{\mathrm{PL}}(i_1, \ldots, i_t, \ldots, i_{K-1}, i_K = 1|\underline{p}) \\
&= \sum_{i_1 \in I\setminus\{1\}} \mathbf{P}_{\mathrm{PL}}(i_1|\underline{p}) \times \cdots \times \sum_{i_t \in I\setminus\{1, i_1, \ldots, i_{t-1}\}} \mathbf{P}_{\mathrm{PL}}(i_t|i_1, \ldots, i_{t-1}, \underline{p}) \times \cdots \times \\
&\times \cdots \times \sum_{i_{K-1} \in I\setminus\{1, i_1, \ldots, i_{K-2}\}} \mathbf{P}_{\mathrm{PL}}(i_{K-1}|i_1, \ldots, i_{K-2}, \underline{p}) \\
&= \sum_{i_1 \in I\setminus\{1\}} \frac{p_{i_1}}{\underline{p}_{[I]}} \times \cdots \times \sum_{i_t \in I\setminus\{1, i_1, \ldots, i_{t-1}\}} \frac{p_{i_t}}{\underline{p}_{[I\setminus\{i_1, \ldots, i_{t-1}\}]}} \times \cdots \times \sum_{i_{K-1} \in I\setminus\{1, i_1, \ldots, i_{K-2}\}} \frac{p_{i_{K-1}}}{\underline{p}_{[I\setminus\{i_1, \ldots, i_{K-2}\}]}} \\
&= \sum_{i_1 \in I\setminus\{1\}} \cdots \sum_{i_t \in I\setminus\{1, i_1, \ldots, i_{t-1}\}} \cdots \sum_{i_{K-1} \in I\setminus\{1, i_1, \ldots, i_{K-2}\}} \frac{p_{i_1} \cdots p_{i_t} \cdots p_{i_{K-1}}}{\underline{p}_{[I]} \cdots \underline{p}_{[I\setminus\{i_1, \ldots, i_{t-1}\}]} \cdots \underline{p}_{[I\setminus\{i_1, \ldots, i_{K-2}\}]}} \\
&= \sum_{i_1 \in I\setminus\{1\}} \cdots \sum_{i_t \in I\setminus\{1, i_1, \ldots, i_{t-1}\}} \cdots \sum_{i_{K-1} \in I\setminus\{1, i_1, \ldots, i_{K-2}\}} \frac{p_{i_1} \cdots p_{i_t} \cdots p_{i_{K-1}} p_1}{\underline{p}_{[I]} \cdots \underline{p}_{[I\setminus\{i_1, \ldots, i_{t-1}\}]} \cdots \underline{p}_{[I\setminus\{i_1, \ldots, i_{K-2}\}]} p_1}.
\end{aligned}$$

The analogous marginal probability corresponding to the selection of an item $l \neq 1$ at stage K is

$$q_l^{[K]} = \sum_{i_1 \in I\setminus\{l\}} \cdots \sum_{i_t \in I\setminus\{l, \ i_1, \ \cdots, \ i_{t-1}\}} \cdots \sum_{i_{K-1} \in I\setminus\{l, \ i_1, \ \cdots, \ i_{K-2}\}} \frac{p_{i_1} \cdots p_{i_t} \cdots p_{i_{K-1}} p_l}{\underline{p}_{[I]} \cdots \underline{p}_{[I\setminus\{i_1, \ldots, i_{t-1}\}]} \cdots \underline{p}_{[I\setminus\{i_1, \ldots, i_{K-2}\}]} \quad p_l}.$$

The $(K-1)!$ ratios in both masses correspond to all possible first $K-1$ stage sampling sequences. We remind the reader that the full-stage sampling sequences $(i_1, i_2, \ldots, i_t, \ldots, i_{K-1}, i_K)$ in the two masses end with 1 and l, respectively. Note that all the $(K-1)!$ ratios that are summed in both expressions have been multiplied respectively by p_1/p_1 and p_l/p_l, so that all the numerators are equal. This simplifies the comparison between $q_1^{[K]}$ and $q_l^{[K]}$ since the numerator is always equal, as in the example of Appendix 1, to the product of all the support parameter components. Hence, if we want to assess the relative magnitude of the two probabilities, we should concentrate on the relative magnitude of the $(K-1)!$ denominators, which all consist of K factors. To this end, one can revisit the notation in order to simplify the comparison task. In the denominators of $q_1^{[K]}$, the K-tuple of indices $(i_1,i_2,\ldots,i_t,\ldots,i_{K-1},i_K)$ is such that the last entry i_K is fixed ($i_K = 1$), whereas the first $K-1$ entries range over the set of permutations of the remaining integers in $A = I\backslash\{1\}$. We can list the set of permutations of the K-tuple of indices $(i_1,i_2,\ldots,i_t,\ldots,i_{K-1},1)$ by using $(a_\sigma,1)$ with $a = (2,3,\ldots,K)$, $a_\sigma = (a_{\sigma(1)},\ldots,a_{\sigma(\mathrm{K}-1)})$ and $\sigma \in S_{K-1}$. Similarly, in the denominators of $q_l^{[K]}$, the last component of the stage sampling sequence is fixed ($i_K = l$), whereas the first $K-1$ entries range over the permutations of the integers in $B = I\backslash\{l\}$. Analogously, we can list the set of permutations of the K-tuple of indices $(i_1,i_2,\ldots,i_t,\ldots,i_{K-1},l)$ by using (b_σ, l) with $b = (2,3,\ldots,l-1,1,l+1,\ldots,K)$, $\mathrm{b}_\sigma = (\mathrm{b}_{\sigma(1)},\ldots,b_{\sigma(K-1)})$. In so doing, we can make a one-to-one comparison of all the homologous denominators, respectively indexed by $(a_\sigma,1)$ and (b_σ,l), by using the same $\sigma \in S_{K-1}$. We remark that, in this way, the $\sigma(t)$th component of a_σ coincides with the $\sigma(t)$th component of b_σ, the only exception being $\sigma(t^*) = l-1$ for which $a_{\sigma(t^*)} = l$ and $b_{\sigma(t^*)} = 1$. In order to rewrite the K factors in the denominators, we will use the subsets A_t^σ and B_t^σ representing the item subsets which comprise, regardless of their order, the components $(a_{\sigma(t)},\ldots,a_{\sigma(K-1)},1)$ and $(b_{\sigma(t)},\ldots, b_{\sigma(K-1)},l)$, respectively. Hence, the homologous denominators to be compared will be written as $\underline{p}_{[A_1^\sigma]} \times \ldots \times \underline{p}_{[A_t^\sigma]} \times \ldots \times \underline{p}_{[A_K^\sigma]}$ and $\underline{p}_{[B_1^\sigma]} \times \ldots \times \underline{p}_{[B_t^\sigma]} \times \ldots \times \underline{p}_{[B_K^\sigma]}$. Now we observe that $A_1^\sigma = B_1^\sigma = I$, hence the first factor $\underline{p}_{[A_1^\sigma]} = \underline{p}_{[B_1^\sigma]} = \underline{p}_{[I]}$ is equal to the sum of all the support parameter components. For similar arguments and in the light of the previous remarks, we can claim that $\underline{p}_{[A_t^\sigma]} = \underline{p}_{[B_t^\sigma]}$ also for $t \leq t^*$. For all $t > t^*$, we have that A_t^σ differs from B_t^σ since, by construction, the former always contains item 1 and not item l, while the latter always contains item l but not item 1, as is apparent by comparing $(a_{\sigma(t)},\ldots, a_{\sigma(K-1)},1)$ and $(b_{\sigma(t)},\ldots,b_{\sigma(K-1)},l)$. Hence, the sums $\underline{p}_{[A_t^\sigma]}$ and $\underline{p}_{[B_t^\sigma]}$ differ only for the presence of p_1 in the former, replaced by the presence of p_l in the latter. Since $p_1 \leq p_l$, this implies $\underline{p}_{[A_t^\sigma]} \leq \underline{p}_{[B_t^\sigma]}$. Finally, for all the K factors, we have $\underline{p}_{[A_t^\sigma]} \leq \underline{p}_{[B_t^\sigma]}$ for all t and σ, hence the opposite inequality holds for the sum of the reciprocals, yielding $q_1^{[K]} \geq q_l^{[K]}$ for $l \neq 1$. The same argument, iterated for each item i such that $p_i \leq p_l$, leads to

$$q_1^{[K]} \geq \cdots \geq q_K^{[K]},$$

that is, the probability masses of the marginal item distribution at the last stage follow the reverse order of the support parameters. We referred to this property as ***inverse monotonicity of the last-stage item probabilities***. The leading argument of the proof is the definition of a one-to-one mapping between the $(K-1)!$ terms of the two sums $q_1^{[K]}$ and $q_l^{[K]}$, obtained by matching the selection stage of the lth item in the sequence $(i_1,i_2,\ldots, i_t,\ldots,i_{K-1},1)$ with the selection stage of item 1 in the sequence $(i_1,i_2,\ldots,i_t,\ldots,i_{K-1},l)$ and all the other item selections in the first $K-1$.

Appendix 3: An example of matrix $\mathbf{T}(\underline{\pi})$ under the EPL specification

By using the rPLMIX function of the *PLMIX* package in R (Mollica & Tardella, 2020), one can simulate $N = 100$ orderings of $K = 5$ items from a genuine EPL model, with a parameter configuration given by

$$\rho = (1,\ 5,\ 2,\ 4,\ 3) \qquad \underline{p} = (0.15,\ 0.4,\ 0.12,\ 0.08,\ 0.25).$$

The code to obtain the synthetic data set of this example is

```
library(PLMIX)
set.seed(12)
sim_data=rPLMIX(n = 100, K = 5, G = 1, p = t(c(0.15,0.4,0.12,0.08,0.25)), ref_order=t(c(1,5,2,4,3)))
```

Under the above EPL specification, the expected rankings of the items in order of occurrence at the first and the last stage are indicated in the two rows of Table 3.

Table 3. Expected rankings of the items in terms of number of selections at the first and the last stage for an $EPL(\rho = (1,\ 5,\ 2,\ 4,\ 3),\ \underline{p} = (0.15,\ 0.4,\ 0.12,\ 0.08,\ 0.25))$ specification. The true first and last stage ranks correspond, respectively, to ranks 1 and 3

	Item				
	1	2	3	4	5
rank$\left(\underline{r}_1^{[1]}\right)$	3	1	4	5	2
rank$\left(\underline{r}_3^{[K]}\right)$	3	5	2	1	4
Sum of ranks	6	6	6	6	6

The matrix $\boldsymbol{T}(\underline{\pi}) = (T_{jj'}(\underline{\pi}))$ for all pairs $j,j' = 1,\ldots,K$ is shown in Table 4. The observed value of the EPL statistic is $T_m(\underline{\pi}) = 0$, which is actually the global minimum of the whole matrix $\boldsymbol{T}(\underline{\pi})$. In this example the global minimum is attained in correspondence of the pair consisting of the true first- and last-stage ranks, namely $(j,j') = (1,3)$.

Table 4. Matrix $\boldsymbol{T}(\underline{\pi})$ for a simulated sample from the $EPL(\rho = (1,\ 5,\ 2,\ 4,\ 3),\ \underline{p} = (0.15,\ 0.4,\ 0.12,\ 0.08,\ 0.25))$. The global minimum is highlighted in bold, as well as the pair in correspondence of which the minimum is attained.

	j'				
j	1	2	**3**	4	5
1	12	8	**0**	6	12
2	8	12	8	10	6
3	0	8	12	10	2
4	6	10	10	12	6
5	12	6	2	6	12

Appendix 4: Estimated Type I error rate and power of alternative goodness-of-fit diagnostics for the EPL assumption Tables 5-9

Table 5. Estimated Type I error rate of alternative goodness-of-fit diagnostics for the EPL assumption

	$K = 5$			$K = 10$			$K = 20$			$K = 40$		
	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$
T_m	0.00	0.00	0.01	0.01	0.04	0.04	0.05	0.04	0.02	0.00	0.01	0.02
X^2_{TOP}	0.00	0.02	0.01	0.01	0.01	0.03	0.03	0.00	0.02	0.22	0.33	0.31
X^2_M	0.00	0.00	0.01	0.00	0.00	0.01	0.03	0.00	0.02	0.47	0.74	0.83
X^2_{PC}	0.01	0.00	0.05	0.02	0.02	0.04	0.02	0.00	0.02	0.13	0.21	0.17
X^2_{IIA}	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.07	0.12

Table 6. Estimated power of alternative diagnostics for the EPL assumption under the DBM-Ken sampling distribution

	$K = 5$			$K = 10$			$K = 20$			$K = 40$		
	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$
T_m	0.01	0.06	0.00	0.04	0.06	0.04	0.24	0.24	0.23	0.81	0.85	0.76
X^2_{TOP}	0.21	0.44	0.65	0.35	0.72	0.61	0.36	0.59	0.61	0.63	0.23	0.93
X^2_M	0.60	0.89	0.97	0.87	0.95	0.93	1.00	0.98	0.99	0.99	1.00	1.00
X^2_{PC}	0.19	0.29	0.63	0.40	0.79	0.81	0.81	0.74	0.85	0.99	1.00	1.00
X^2_{IIA}	0.52	0.88	0.95	0.82	0.93	0.91	0.95	0.97	0.99	0.97	1.00	1.00

Table 7. Estimated power of alternative diagnostics for the EPL assumption under the DBM-Cay sampling distribution

	$K = 5$			$K = 10$			$K = 20$			$K = 40$		
	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$
T_m	0.44	0.57	0.48	0.65	0.66	0.72	0.56	0.66	0.77	0.56	0.55	0.60
X^2_{TOP}	0.78	0.89	0.85	0.81	0.80	0.84	0.70	0.75	0.87	0.69	0.69	0.74
X^2_M	0.85	0.90	0.90	0.84	0.83	0.89	0.82	0.79	0.89	0.77	0.80	0.81
X^2_{PC}	0.78	0.85	0.88	0.83	0.79	0.83	0.70	0.75	0.85	0.66	0.68	0.72
X^2_{IIA}	0.76	0.81	0.84	0.74	0.71	0.79	0.56	0.68	0.74	0.49	0.55	0.58

Table 8. Estimated power of alternative diagnostics for the EPL assumption under the DBM-Ham sampling distribution

	$K = 5$			$K = 10$			$K = 20$			$K = 40$		
	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$
T_m	0.26	0.38	0.42	0.72	0.78	0.74	0.61	0.74	0.79	0.47	0.61	0.54
X^2_{TOP}	0.82	0.86	0.89	0.84	0.95	0.91	0.73	0.79	0.85	0.71	0.70	0.70
X^2_M	0.87	0.89	0.93	0.91	0.95	0.93	0.82	0.80	0.88	0.82	0.80	0.78
X^2_{PC}	0.83	0.86	0.90	0.85	0.94	0.89	0.76	0.77	0.86	0.70	0.71	0.69
X^2_{IIA}	0.81	0.86	0.85	0.79	0.91	0.87	0.61	0.70	0.77	0.55	0.59	0.57

Table 9. Estimated power of alternative diagnostics for the EPL assumption under the TH-norm sampling distribution

	$K = 5$			$K = 10$			$K = 20$			$K = 40$		
	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$	$N = 300$	$N = 450$	$N = 600$
T_m	0.08	0.13	0.18	0.46	0.42	0.43	0.64	0.79	0.85	0.80	0.83	0.87
X^2_{TOP}	0.63	0.74	0.90	0.83	0.96	0.92	0.92	0.94	0.97	0.98	1.00	1.00
X^2_M	0.94	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X^2_{PC}	0.64	0.76	0.88	0.92	0.95	0.99	0.94	0.99	0.98	0.98	1.00	1.00
X^2_{IIA}	0.76	0.87	0.92	0.94	0.98	1.00	0.98	1.00	1.00	1.00	1.00	1.00