Investigating the Judges Performance in a National Competition of Sport Dance

Noname manuscript No.
(will be inserted by the editor)

Investigating the judges' performance in a national competition of sport dance

Laura Anderlucci · Alessandro Lubisco · Stefania Mignani

Received: date / Accepted: date

Abstract Many sports, such as gymnastics, diving, figure skating, etc. use judges' scores to generate a rank for determining the winner of a competition. These judges use some type of rating scale when assessing performances. Human ratings are subject to various forms of error and bias. The overall outcomes may largely depend upon the set of chosen raters. The aim of this paper is to illustrate how results from the Many-Facet Rasch Measurement framework (MFRM) can be used to highlight feedback to judges about their scoring patterns. The purpose is to analytically detect anomalous rater behaviours. We consider the field of Sport Dance, a discipline which enjoys increasing public interest and passion in recent years. We analyze data relating to two national competitions held in Italy in 2018 and 2019.

Keywords Many-Facet Rasch Measurement · rater effect · aesthetic sport

1 Introduction

"Who judges a judge?" is a delicate question that offers food for thought in different fields: from education to justice, from entertainment to sport, just to name a few. Each organized system needs rules and regulations in order to work in a proper and coherent way. For decades, psychological research has investigated the subjective component in the rating process of many fields,

L. Anderlucci
University of Bologna, via Belle Arti 41, 40126 Bologna, Italy
Tel.: +39 051 2098267
E-mail: laura.anderlucci@unibo.it

A. Lubisco
University of Bologna, via Belle Arti 41, 40126 Bologna, Italy

S. Mignani
University of Bologna, via Belle Arti 41, 40126 Bologna, Italy

trying to validate through statistics such processes as proper measuring tools (Saal et al., 1980).

Within the evaluation of sport performance field, two main questions can rise: the first is about the criteria determining the assignment of a score, the second pertains to the objectivity and impartiality of the judgements. If the former question can be addressed by consultation of specific regulations redacted by sport federations, the latter is more problematic and represents the aim of the present work.

Competition judges are professionals in specific disciplines; very often, they are former athletes and trainers, therefore they have deep knowledge of the technical aspects and the necessary experience to evaluate athlete performances. In Italy, the title of competition judge can be obtained only after passing a professional examination at the end of a dedicated education program. The education program aims at enabling judges to evaluate both technical and non-technical aspects of a particular activity.

However, the education does not guarantee that the evaluation process is cleaned of any sort of conditioning or of mistakes that interfere with proper judgment. It is a multiplicity of distortion factors, mainly tied up to psychological conditioning, that takes action during an evaluation process. It is not necessary to indicate the bad intentions of the judge to understand that there are many sports in which the subjective interpretation leads to questionable decisions, even when these are made within the regulations. The competitive aspect of the modern sport may conflict with different kinds of interest. For this reason, it is essential to have a system of refereeing and evaluation that is as free from subjective considerations as possible.

For example, the judging system of figure skating changed in 2004, as a consequence of the 2002 Winter Olympics scandal, whose controversy led two teams receiving the gold medal. The so-called "6.0 system", where each judge would rank the athletes, has been replaced by the International Judging System, where the base value of each element performed by the skater is identified by a Technical Panel and then graded by a Judging Panel. The episode has underlined the importance to restrict the rater's bias and to make the scoring system more objective and less vulnerable to abuse.

This paper is an attempt to analyze this kind of problem in an unusual area: sport dance. Sport dance is based on a rating system that will be described later, in Section 4.

We focus on two datasets from the Italian National Competition of Sport Dance that takes place in Rimini (Italy) every year, at the beginning of the summer; specifically we consider the category Synchro Latin, under 11 of 2018 and under 15 of 2019. The scores of athletes were analysed with the purpose of providing in-depth feedback to judges about their scoring patterns.

The collected data have a typical three-way structure: dancers (specifically, teams of dancers), scores on four dancing features, and occasions represented by the judges. Several methods for multi-way data have been proposed in the literature, for example Parallel Factor Analyis (PARAFAC) or Tucker3 (Carroll and Chang, 1970; Harshman, 1970; Tucker, 1966). These methods

decompose the original data array into sets of scores and loadings with the idea of describing the data in a more condensed form.

In this paper we considered another possible approach in the context of Rasch models, with the aim of estimating a measurement model that provides a fine-grained analysis of multiple variables potentially having an impact on assessment outcomes. In general, the Factor Analysis approach is more flexible when the level of modelling abstraction is focused on the relationship within a system of variables; differently, the Rasch approach is generally preferred when the items are the focus of interest and the objective is to create a unidimensional score scale that summarizes the items as a whole, without modelling underlying concepts. In particular, we applied the Many-Facet Rasch Measurement (MFRM) models (Linacre, 1989), that extend the basic Rasch model (Rasch, 1960; Wright and Stone, 1979) to incorporate more elements (i.e. *facets*) than just two.

In general, ability assessments are characterized by distinct sets of aspects, directly or indirectly involved in determining the measurement outcomes. In this context a facet can be defined as any factor, variable, or component of the measurement process that is assumed to affect scores in a systematic way (Bachman, 2004; Linacre, 2002; Wolfe and Dobria, 2008). This definition includes facets that have key roles in the analysis (e.g. examinees, items or tasks), and facets that are assumed to contribute to systematical measurement error (e.g. raters, interviewers, time of testing). Moreover, facets can interact with each other in various ways.

In this paper we compare the estimated ranking returned by MFRM (and based on raw scores assigned by the judges) with the actual ranking, provided by the committee after the application of a specific adjustment, as described in Section 4. Such comparison allows highlighting differences in what to expect and what is observed, and it strengthens the need for a proper monitor on the raw scores, so as to obtain a more reliable result. In fact, the two competitions under study are examples of scenarios where judges behave very differently when assigning scores: the competition of 2018 shows results similar to the theoretical ones resulting from the model, whereas results from the second competition identify a much stronger rater effect.

The paper is organized as follows. Section 2 provides an overview of the rater effect in evaluation procedures. Section 3 briefly describes the Many-Facet Rasch Measurement models and the fit indices that will be studied in the data analysis section; Section 4 explains the rules of Sport Dance competitions and their rating systems. Data are introduced and the results are presented in Section 5. A brief discussion concludes the paper.

2 The rater effect in the evaluation procedure

The reliability and validity of performance ratings have long been a source of concern to researchers and practitioners alike. When talking of *rater bias* we refer to the variability in the scores not due to the ability of the athletes

but to the characteristics of the raters. This source of variability could compromise the fairness of the judgments and it is called "rater effect" (Parke et al., 2006; Roever and McNamara, 2006; Murphy and Balzer, 1989). Rater variability manifests itself in several ways, each deserving close investigation. Rater effects, often discussed in the literature, can be distinguished in *severity/leniency*, *halo*, and *central tendency* effects.

The severity effect occurs when raters provide ratings that are, on average, lower than those assigned by other raters, even after accounting for the ratee performances. In other words, there is severity when a ratee is given a score lower than his/her behaviour or ability would actually deserve. Leniency works in the opposite direction, and consists in assigning a higher score compared to ability.

The central tendency is the condition where raters avoid the extreme categories of a rating scale and overuse the middle ones. The raters show inability to differentiate among ratee performance levels along the entire performance continuum. In other words, raters do not understand the distinctions between any of the scale categories, and thus resort to assigning all ratees similar "middle-of-the-road" ratings. Central tendency represents a special case of the so-called *range restriction*, that is a restricted use of the available values, not necessarily around the central points of the scale.

The Halo effect is defined as rater's tendency to give the same score to distinct features of ratee performance, being influenced by overall impression of a given performance or by a single feature viewed as highly important.

3 The methodology for the analysis of rater effect

In this paper the Many-Facet Rasch Measurement (MFRM) approach has been used to detect and to measure rater effects (see e.g. Engelhard, 2002; Knoch, 2009; Linacre, 2009; Myford and Wolfe, 2003, 2004). Within the framework of sport evaluation, we can define the athletic competition as the evaluation test, whose item responses are represented by the scores of specific elements of judgment. Therefore, teams of athletes can be seen as ratees and their skills as global performance.

The MFRM model has all of the characteristics of basic Rasch model and also has other advantages. It is characterized by specific objectivity (or relational invariance), linearity, and measurement units. The measures obtained by the model are sample-, item-, and condition-free. The specified facets are analyzed simultaneously and calibrated onto a single linear scale (i.e. the logit scale). This property allows comparisons between facets. In addition, it is possible to analyse with more details individual level effect within each facet; for instance, which raters judge more severely or which raters disagree about with other raters.

Furthermore, useful tools to assess the fit for each element of each facet are available; such indices measure the degree of similarity between observed ratings and the corresponding expected ones estimated by the model. Therefore,

fit indices are useful for detecting various rater effects like severity/leniency and also central tendency or halo effects (Engelhard, 2002; Knoch et al., 2007; Myford and Wolfe, 2003, 2004). Such information completes the picture and represents an important tool in detecting potential anomalies.

3.1 Model Specification

The method used in this work belongs to the family of models defined by that of Rasch. Such models transform dichotomous or ordinal observations into linear measurements through a logistic regression, linking the results obtained in a particular test with the skills of the individuals and the difficulty of the task (Farrokhi and Esfandiari, 2011). Within the framework of sport evaluation we can define the test as the competition and the items with the scores of specific elements of judgment. Therefore, athletes can be seen as ratees and their skills as global performance.

When items are dichotomous the methodology resorts to the classical Rasch model; differently, in presence of ordinal data a possible option is provided by the Rating Scale Model (Andrich, 1978), that is the one applied in this work.

Indicating with n the ratee and with i the item that can assume K ordinal values ($k = 1, \ldots, K$), the model has the following structure:

$$\log \frac{P_{nik}}{P_{ni(k-1)}} = B_n – D_i – F_k \tag{1}$$

where P_{nik} indicates the probability of observing category k of the i-th item on ratee n, while $P_{ni(k-1)}$ is the equivalent for category $k-1$, B_n is the skill of ratee n, D_i is the difficulty of item i and F_k the difficulty of scale category k relative to scale category k–1 for all of the items.

Linacre (1994) introduces a third element related to the judges, so as to estimate their strictness. Including such element, model (1) becomes the Many-Facet Rasch Measurement. The model in its basic formulation can therefore be written as (Myford and Wolfe, 2003):

$$\log \frac{P_{nijk}}{P_{nij(k-1)}} = B_n – D_i – C_j – F_k \tag{2}$$

where C_j is the severity of rater j, that assigns score k for the item i on the ratee n.

When the model is estimated, all the facets are analyzed simultaneously but independently. This means that the estimate of any parameter is dependent on the accumulation of all ratings in which it participates, but is independent of the particular values of any of those ratings. This axiom of "local independence" allows the statistical estimates of the measures to be as free as possible from which particular judge rated which particular examinee on which particular item. Facets are calibrated on the single linear scale determined by the logit; such joint calibration permits measurement on the same scale, i.e.

to compare, the strictness of the rater, the performance of the ratee and the difficulty of the item (Myford and Wolfe, 2003).

The model is estimated via Joint Maximum Likelihood, with no restrictions on the parameters; the software we used is Facet, developed by Linacre (Linacre, 2013).

3.2 Goodness-of-fit

Rasch models are idealizations of empirical observations. Therefore, empirical data will never fit a given Rasch model perfectly (Eckes, 2011). Assessing the global fit of data to a model may thus translate into a futile endeavour. Hence, a reasonable strategy is to explore the practical utility or significance of a model. In other words, it is essential to know whether the data fit the model usefully, and, when misfit is found, it is important to evaluate its magnitude, its origin and decide what to do about it. Measures of fit enable detection of potential errors in the evaluation system and will be briefly discussed in the following.

A pivotal measure of fit is provided by the residuals. Residuals may indicate the degree to which observed ratings match the expected ratings that are generated by the Many-Facet Rasch model:

$$R_{nij} = X_{nij} - E_{nij},$$

where $E_{nij} = \sum_{k=0}^{K} kP_{nijk}$ denotes the expected rating, based on Rasch parameter estimates for ratee n on feature i by rater j, and X_{nij} the corresponding observed rating. Large differences between the observed and expected scores, particularly for individual raters, may suggest the existence of rater effects.

In order to account for individual variability, residuals are then standardized:

$$Z_{nij} = \frac{X_{nij} - E_{nij}}{\sqrt{W_{nij}}},$$

where $W_{nij} = \sum_{k=0}^{K}(k - E_{nij})^2 P_{nijk}$ indicates the expected variation of the observed score X_{nij} around its expectation under Rasch-model conditions and is called *model variance*.

Standardized residuals with absolute values greater than 2 have $p < 0.05$ under Rasch-model conditions, indicating ratings that are highly unexpected; those observations may be subjected to closer inspection.

Global fit indices, called *mean-square* (MS), for the elements of each facet can be derived from the average of squared standardized residuals (Linacre, 1994); here, we will focus on the rater fit statistics.

According to the average measure, two statistics can be defined: if the average is *unweighted* the MS is called *outfit*, whereas if it is weighted by the

model variance is named *infit*. The former is simply the average of the rater's squared standardized residuals across all ratees and features rated by that rater. Such measure is sensitive to outlying individual unexpected rates, like a very high score assigned by a severe judge to a poor performer with respect to a hard task.

The latter is sensitive to inlying unexpected ratings. More specifically, infit is sensitive to unexpected ratings where the locations of rater j and the other elements involved are aligned with each other, that is, where the locations are closer together on the measurement scale. In addition, as the weights indicate the amount of statistical information about the elements in question, the infit is also said to be sensitive to unexpected ratings that provide more information. Since such ratings are generally associated with higher estimation precision, infit is commonly considered more important than outfit in judging rater fit (Myford and Wolfe, 2003; Eckes, 2011).

The infit and outfit mean-square indices have an expected value of 1 and can range from 0 to infinity. For both indices, values close to 1 indicate that the observed ratings are close to their expected ratings. Values smaller than one indicate potential overfitting; values larger than 1 may suggest misfit, that is a larger variability than that allowed by the model (Looney, 2004).

Usually, acceptable values lie in an indicative range that spans between 0.6 and 1.4. However, if the results from the analyses are meant to inform high-stakes decision making, Myford and Wolfe (2003) suggest setting more stringent upper- and lower-control limits (for example, an upper-control limit of 1.2 and a lower-control limit of 0.8).

Another interesting measure to look at is the result of the chi-square test for the 'fixed effect' hypothesis; such test allows to determine whether all raters exercised the same level of severity when evaluating ratees, after accounting for measurement error. If the hypothesis is rejected, it is possible to conclude that there are at least two raters that differ. The fixed-effect test can be coherently carried out for the other facets as well.

In addition, the 'Single Rater/Rest Of the Raters' correlation (SR/ROR) carries important information. The SR/ROR correlation (also known as *point-biserial correlation* in Facets documentation) summarizes the degree to which a particular rater's ratings are consistent with the ratings of the rest of the raters. Values of correlations near zero or negative for a given rater indicate that the assigned ranks order ratees in a manner which differs from the other raters' rank ordering.

Another measure that will be used to identify potential rater effect is the *separation statistics*. The rater separation ratio, G, is a measure of the spread of the rater performance measures relative to their precision (Myford and Wolfe, 2003). The rater separation index (H) is obtained as $H = (4G + 1)/3$ and indicates the number of separated homogenous group on raters; for example, a rater separation index of 3.71 suggests that there are about four statistically distinct strata of rater severity.

Finally, the *reliability* of the rater separation index provides information about how well the raters are separated in order to reliably define the rater

facet. It is a measure of the spread of the rater severity measures relative to their precision, reflecting potentially unwanted variation between raters in the levels of severity exercised. In the evaluation context, the most desirable result is to have a reliability of rater separation close to zero, which would suggest that the raters were interchangeable, exercising very similar levels of severity.

4 Sport Dance competition and rating systems

Sport dance represents the trait d'union between artistic expression and competitive approach. The athletes are divided into categories based both on age and skill level. Sport dance includes more than 50 different dancing disciplines divided into two main groups. The first group, i.e. team dances, includes international, national and regional dances, like Viennese Waltz, Samba, Rock'n'Roll, Tango, Mazurka, to name a few. The second group, artistic dances, includes academic, choreographic and street dances, e.g. Modern Jazz, Show Dance, Disco Dance, Break Dance, Hip Hop...

Both in national and international competitions, depending on the disciplines, judges use various methods to evaluate the athlete performance. With the Cross-Skating *comparative system*, depending on the phase of the competition, each judge is asked to:

- point out which athletes in his/her opinion must move on to the next round (X system used in preliminary phases);
- draw up his/her own ranking from 1 to n, where n is the number of athletes (Skating System used in final stages).

Differently, within the *absolute system*, each athlete is given a unique and distinct value. Usually judges use integer values in a 1 to 10 scale, where 1 is the lowest and 10 the highest score. Athletes with the best values in the preliminary phases will pass and move to the subsequent round. In the final phase, the ranking is defined by introducing the calculation of median values in order to limit the effect of outliers, i.e. scores given by judges who want to favour or penalize one or more athletes. All the scores are processed with a specific software that returns the rankings and publishes them on the Web. Further details will be provided later in this Section.

The FIDS (Italian Federation of Sport Dance) uses the so called 3D and 4D international measurement systems. The two systems are identical except for the number of the performance components to be evaluated. In the 3D system judges express their opinion on a 1 to 10 scale for each of the following components: Technique, Choreography and Image, defined respectively with T, C and I. Each athlete may receive a total score from 3 to 30 from each judge. In the 4D system, judges must evaluate also a Show component (S), that is expressed as Synchro for Synchro Dance and Choreographic dance or as Acrobatic for Acrobatic Rock'n'Roll.

In order to better explain how the rating system works, let's consider a dataset from an Under 15 Synchro Latin competition with 4D; such competition evaluates the performances of teams of dancers.

Pos.	Athl.	Phase	C	E	G	I	K	L	M	N	U	Total
1.	(701)	F	8|8|8|9||5	10|9|9|9||4	6|6|7|6||8	7|8|8|7||6	7|8|8|8||2	10|10|9|9||1	9|9|9|9||1	7|8|9|9||2	9|9|9|8||1	1.0
		1	8|8|9|8||x	9|9|9|9||x	7|7|7|7||x	7|8|8|7||-	8|8|8|8||x	10|9|9|9||x	9|9|9|9||x	7|8|9|9||x	8|9|8|8||x	298/8
2.	(708)	F	9|9|8|8||3	6|6|6|6||10	6|6|6|6||9	9|9|8|7||1	8|8|7|7||3	8|8|8|9||6	8|8|8|8||3	7|8|8|9||3	7|8|8|8||11	2.0
		1	8|9|8|8||x	6|6|6|6||-	7|7|8|7||x	9|9|8|8||x	8|8|8|7||x	10|10|10|10||x	8|8|9|9||x	7|7|8|7||x	7|7|8|8||-	284/7
3.	(589)	F	9|10|10|10||1	10|10|10|9||2	7|8|8|7||4	8|7|7|6||9	6|5|5|5||12	7|7|7|7||11	8|9|9|8||2	7|8|9|9||1	8|8|8|8||7	3.0
		1	9|9|9|9||x	10|10|10|9||x	7|7|7|8||x	8|8|7|7||-	9|8|8|8||x	8|8|8|7||-	9|8|9|9||x	8|7|7|8||x	8|8|8|8||x	295/7
4.	(774)	F	8|8|8|9||6	4|4|4|4||12	5|5|6|6||11	8|8|9|7||4	9|9|9|9||1	8|9|9|9||4	7|8|7|7||7	8|7|8|7||4	8|9|8|8||4	4.0
		1	9|8|8|9||x	5|5|5|5||-	6|6|6|6||-	8|9|9|8||x	8|8|8|8||x	7|9|9|8||-	6|7|7|6||-	7|7|8|7||x	8|8|8|8||-	264/4
5.	(706)	F	5|4|5|5||12	10|10|9|9||3	8|8|8|8||2	9|8|8|8||2	7|7|7|7||5	8|8|8|8||7	6|4|4|6||12	6|7|8|8||5	8|8|8|8||8	5.0
		1	4|5|5|6||-	9|9|10|10||x	7|8|8|7||x	9|8|9|9||x	8|8|8|9||x	9|8|8|9||x	6|6|6|6||-	6|5|7|7||-	8|8|8|8||x	271/6
6.	(799)	F	8|8|9|8||4	8|8|8|9||6	5|6|6|6||10	8|8|8|8||5	6|7|7|7||6	9|9|10|9||2	6|7|7|7||9	7|7|7|7||6	8|8|9|9||3	6.0
		1	8|8|9|8||x	7|7|8|8||-	4|5|6|6||-	8|9|9|8||x	7|8|7|8||-	9|10|10|9||x	6|6|7|6||-	7|6|6|6||-	8|8|9|9||x	270/4
7.	(806)	F	6|6|6|7||8	8|8|8|8||7	7|7|7|7||6	7|7|8|8||7	6|7|7|6||7	9|9|9|9||3	8|8|7|8||4	5|5|6|6||12	7|8|8|8||12	7.0
		1	7|8|7|7||x	5|5|5|5||-	5|5|5|5||-	8|8|8|8||x	7|7|7|7||-	10|9|10|10||x	9|7|8|8||x	6|6|6|6||-	8|8|8|8||-	256/4
8.	(670)	F	5|4|5|6||11	9|9|9|9||5	8|7|8|7||3	8|8|9|8||3	6|5|6|6||10	7|8|7|8||9	6|6|6|6||11	6|6|8|8||7	8|9|8|8||5	8.0
		1	7|7|6|7||-	10|9|9|9||x	7|7|7|8||x	8|8|8|8||x	8|7|7|7||-	8|8|8|8||-	7|7|6|6||-	6|6|7|7||x	7|9|8|8||-	270/4
9.	(724)	F	8|8|7|5||7	5|5|5|5||11	4|5|5|5||12	9|8|8|1||12	6|7|7|5||8	8|8|9|9||5	6|6|7|7||10	7|7|6|7||8	9|9|8|8||2	9.0
		1	8|8|8|8||x	6|6|6|6||-	5|5|6|5||-	9|8|8|9||x	7|7|7|8||-	10|10|10|10||x	6|6|7|7||-	7|7|7|7||x	9|9|8|8||x	268/5
10.	(715)	F	5|4|7|7||9	7|7|7|7||8	7|7|8|8||5	8|7|8|6||8	6|6|6|6||9	7|7|8|9||8	8|7|7|7||6	5|5|6|6||11	7|8|9|8||9	10.0
		1	7|7|7|8||-	7|7|7|7||-	7|7|7|7||x	8|8|8|8||x	8|7|8|8||x	10|9|9|10||x	7|8|7|7||x	6|6|6|7||-	7|8|9|8||-	272/5
11.	(588)	F	5|5|6|5||10	10|10|10|10||1	8|8|8|9||1	7|7|7|7||10	6|7|8|8||4	8|7|7|7||10	7|8|6|7||8	7|6|7|7||9	7|8|8|8||10	11.0
		1	5|3|5|4||-	10|10|10|10||x	7|7|7|7||x	7|8|7|7||-	7|8|8|8||x	8|8|9|9||-	8|8|8|8||x	6|7|7|7||x	8|8|8|8||x	270/6
12.	(622)	F	9|9|9|9||2	6|6|6|6||9	7|6|7|6||7	7|7|7|6||11	6|6|5|5||11	7|7|6|7||12	8|8|7|7||5	6|6|7|7||10	7|9|9|8||6	12.0
		1	7|7|7|7||-	7|7|7|6||-	6|7|7|6||-	7|7|7|6||-	8|8|8|8||x	7|7|6|6||-	7|8|8|7||x	6|7|7|6||x	8|8|8|8||x	254/4
13.	(674)	1	6|7|6|7||-	8|8|8|8||x	7|7|8|8||x	8|9|8|7||x	6|7|8|7||-	8|8|7|7||-	5|6|6|6||-	4|5|6|8||-	7|8|8|8||-	255/3
14.- 15.	(689)	1	5|5|6|7||-	10|10|9|9||x	7|7|6|6||-	7|7|7|7||-	5|7|7|6||-	6|5|6|6||-	6|8|8|8||x	6|6|8|7||x	7|8|8|8||-	251/3
14.- 15.	(748)	1	7|7|7|7||-	7|7|7|7||-	7|8|7|8||x	7|7|7|5||-	6|7|7|6||-	6|7|7|7||-	8|6|7|7||x	7|7|6|7||x	8|7|8|8||-	251/3

Fig. 1 Individual scores and final ranking of the Under 15 Synchro Latin competition.

Figure 1 displays the ranking after the final round; each row refers to an athlete and the rows are ordered by the final ranking shown in the "Pos" column (that stands for 'Position'). In particular, in this case, each row represents a dance team ("Athl." column) identified by a number: 701, 708, 589.... Names are omitted. In column "Phase", *F* stands for *Final stage*, while 1 means *Preliminary phase*. Each column includes scores assigned by each of the judges, nine in this case, indicated with capital letters C, E, G,..., U. The last column, "Total", contains the final rank of the first 12 teams only and the total score of the preliminary phase for all the teams.

For example, in the preliminary phase team 674 in 13th position received from Judge C the scores 6, 7, 6 and 7 for Technique, Choreography, Image and Synchro, respectively. The same team received four 8 from Judge E. And Judge E indicated with an *X* such team 674, meaning that, in his/her opinion, it was one of the best 12 (following the comparative *X* System). Unfortunately, only three judges (E, G and I) gave an *X* to team 674 and, for this reason, it was out of the final. In fact, all the other teams that passed the preliminary phase received a larger number of *X*s.

We shall now focus on team 701, who won the competition: they received 8, 8, 8 and 9 from Judge C. This is the fifth highest score for this judge; in other words, team 701 scored 5th for Judge C. Only judges L and U put team 701 in first position. So, how is the final ranking obtained?

Figure 2 shows the intermediate evaluations that led to the final ranking.

In this phase only the rankings of each judge are considered. Team 701, the winner of the competition, is 5th for Judge C, 4th for Judge E, 8th for Judge G, and so on. The ordered sequence of the rankings is 1, 1, 1, 2, 2, 4, 5, 6, 8. This means that the median ranking expressed by the nine judges is 2. This is shown by the value of 5 in column "1.-2." that means: there are 5 judges

Athl.	Judges									Calculation											Pos.
	C	E	G	I	K	L	M	N	U	1.	1.-2.	1.-3.	1.-4.	1.-5.	1.-6.	1.-7.	1.-8.	1.-9.	1.-10.	1.-11.	1.-12.	
	Positions																				
588	10	1	1	10	4	10	8	9	10	2	2	2	3	3	3	3	4	5(23)	-	-	-	11,0
589	1	2	4	9	12	11	2	1	7	2	4	4	5(10)	-	-	-	-	-	-	-	-	3,0
622	2	9	7	11	11	12	5	10	6	0	1	1	1	2	3	4	4	5(29)	-	-	-	12,0
670	11	5	3	3	10	9	11	7	5	0	0	2	2	4	4	5	-	-	-	-	-	8,0
701	5	4	8	6	2	1	1	2	1	3	5	-	-	-	-	-	-	-	-	-	-	1,0
706	12	3	2	2	5	7	12	5	8	0	2	3	3	5	-	-	-	-	-	-	-	5,0
708	3	10	9	1	3	6	3	3	11	1	1	5	-	-	-	-	-	-	-	-	-	2,0
715	9	8	5	8	9	8	6	11	9	0	0	0	0	1	2	2	5(35)	-	-	-	-	10,0
724	7	11	12	12	8	5	10	8	2	0	1	1	1	2	2	3	5(30)	-	-	-	-	9,0
774	6	12	11	4	1	4	7	4	4	1	1	1	5(17)	-	-	-	-	-	-	-	-	4,0
799	4	6	10	5	6	2	9	6	3	0	1	2	3	4	7	-	-	-	-	-	-	6,0
806	8	7	6	7	7	3	4	12	12	0	0	1	2	2	3	6	-	-	-	-	-	7,0

Fig. 2 Intermediate steps to produce the final ranking of the Under 15 Synchro Latin competition.

that put that team in a position from 1 to 2. No other team has median values lower than this. For this reason, team 701 is the winner.

Team 708, is 2nd in the final ranking because the median value of its single rankings is 3, the second best. But, how is the 3rd position determined? Two teams, 589 and 774, have a median value of their rankings equal to 4. In this case, the software calculates the sum of the first 5 rankings in the ordered list. The sequence for team 589 is 1, 1, 2, 2, 4, 7, 9, 11, 12. As has already been said, the median is 4 and the sum of the first 5 values is 10. The column "1.-4." contains the value "5(10)", which stands for "Median=4" and "Sum of the first 5 values=10". For team 774 the ordered sequence is 1, 4, 4, 4, 4, 6, 7, 11, 12. Same median as team 589, but the sum of the first 5 ranks is 17. This means that the first five rankings of team 589 are better than those of team 774. Team 589 was classified 3rd and team 774 4th.

Using the median of the rankings given by each judge rather than the overall score to determine the final ranking has two main benefits. First of all, it accounts for different leniency/severity: using the ranks rather than raw values is more robust. Secondly, considering the median rank prevent from potential favouritism or partiality, as they would have an effect only on the tails of the rank distribution.

5 Analysis of the raters' behaviour

5.1 Description of the Datasets

Data come from the Italian National Championship of SportDance, held in Rimini (Italy) in July 2018 and July 2019. All the results are available from the website of the Italian Federation of SportDance. Results are distinct for competition site and date.

The first dataset[1] we analyzed refers to the Synchro Latin Dance discipline competition, disputed on July 5th 2018 at the Arena Bianca, group U/11C (Under 11). Such competition included 14 teams and 9 raters.

[1] Data available at the website `https://www.federdanza.it/images/gare/2017_2018/EXPORT/CAMPIONATI_2018/arena_bianca_05/12-grp-d_synchrolat_u11_c/index.htm`

The second dataset[2] pertains to the Synchro Latin Dance discipline as well, competition disputed on July 5th 2019 at the Arena Bianca, group U/15C (Under 15). Such competition included 12 teams and 9 raters.

In both cases, each team was evaluated according to four features: Technique, Choreography, Image and Show, all measured on a 10 point-scale. Judges are indicated with capital letters, the athletes via a number that corresponds to the position in the actual final ranking.

5.2 Results of the Many-Facet Rasch Measurement model

We considered the model of Equation (2) and it was estimated via *Facet* software (Linacre, 2013).

A logit estimate of the calibration with the corresponding standard error for each element of each facet are provided. Figures 3 and 4 contain the ruler (also called *variable map*) that displays all the considered facets on a single scale.

5.2.1 Results of Synchro Latin Dance, group U/11C

Figure 3 contains the variable map of the estimation results.

The first column of the ruler displays the linear, equal-interval logit scale upon which all facets in the analysis are positioned, determining a unique reference for comparisons within and between the facets. It can be seen that the logit measure spans from -2 to +3.

The second column displays the nine rater severity measures, obtained by the inclusion of parameter C_j in Equation (2). The judges are ordered in terms of leniency/severity each rater exercised when evaluating the athletes. Since the facet 'Judge' was included as positively correlated with the logit score, more severe raters appear at the bottom of the column, whereas more lenient raters appear at the top. For the Under 11 competition it is possible to identify two judges that were more lenient (i.e. Judge Q and Judge I) and two judges that behaved more severely (namely, Judges P and M) compared to the others; their behaviour does not lie on the very extreme points of the scale, but still quite far from a more 'neutral' score around zero.

The third column displays the four feature difficulty measures. These measures are produced as a result of including the D_i parameter in Equation (2). Features appearing lower in the scale were more difficult than those appearing higher. Here, the four aspects do not differentiate much in terms of difficulty; Image seems to be the feature more likely to obtain higher scores.

The fourth column includes the performance measures of the athletes. These measures are produced as a result of including the B_n parameter in model (2). Ratee performance measures are single number summaries on the logit scale of each ratee's tendency to receive high or low ratings across raters

[2] Data available at the website `https://www.federdanza.it/images/gare/2018_2019/EXPORT/arena_bianca_05/10-p-grp_synchrolat_u15_c/index.htm`

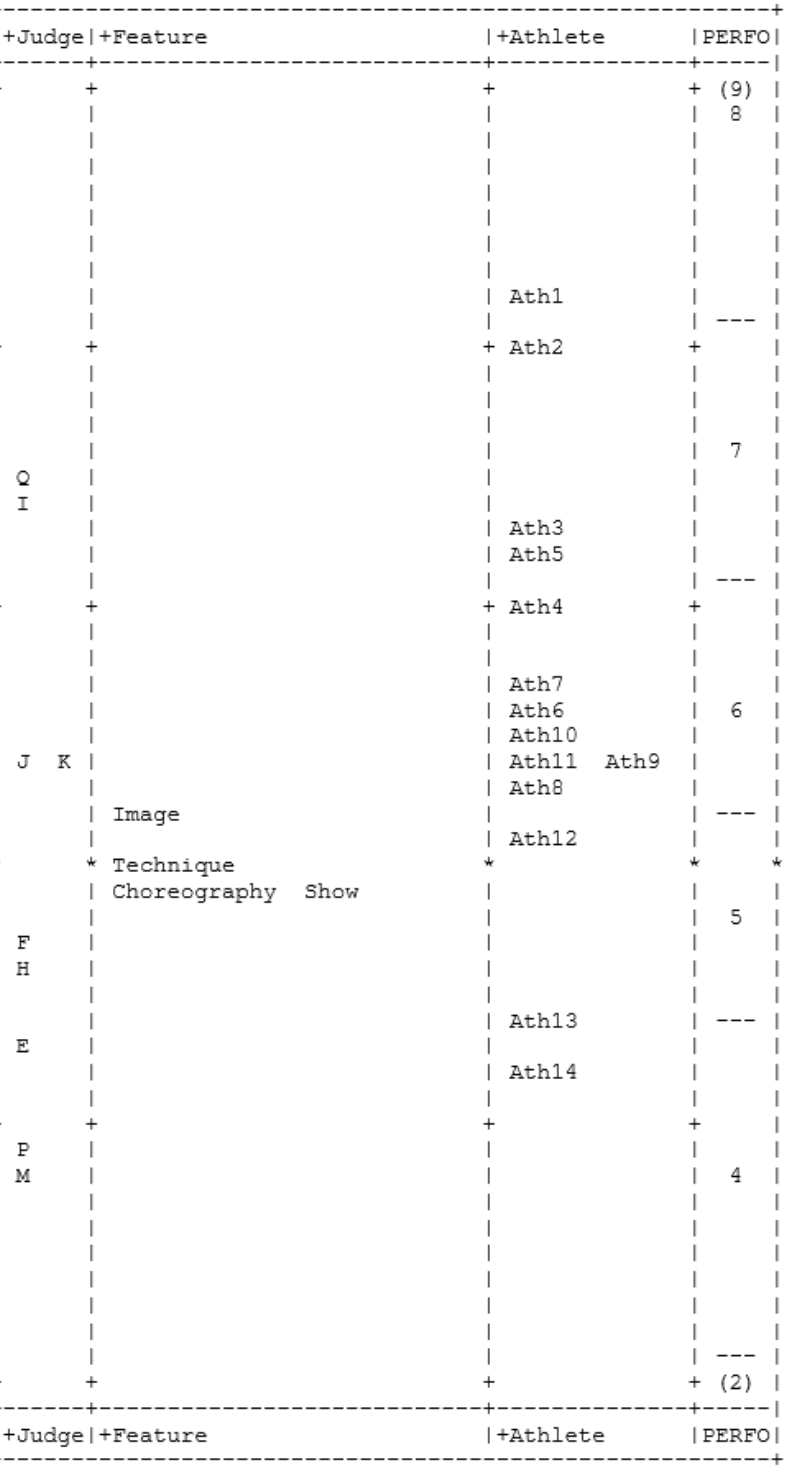

Fig. 3 Variable Map from Facets Analysis of the first dataset [U/11C in 2018], with three facets: athletes, aspects, and judges (raters).

on the four features. The ratees are ordered from highest performing (at the top of the column) to lowest performing (at the bottom of the column). As the athletes are named according to the final ranking, it can be noticed that the podium and the last positions are indeed placed at the top and at the bottom of the scale; central positions are not perfectly reproduced by the model, suggesting that such performances were harder to evaluate in a objective and coherent way.

The last column displays the supposed ten-point rating scale. 'Supposed' since the scale actually ranges between 2 and 9, rather than from 1 to 10; in addition, score 3 was never assigned. The horizontal lines across the column indicate the threshold above which the likelihood of a ratee to receive the next higher score begins to exceed the likelihood of that ratee to receive the next lower score. For example, athletes with performance measures between 1 logits and 2 logits are more likely to receive a rating of 7 than any other rating across the four features.

Table 1 Judge Measurement Report of the first dataset, U/11C in 2018

Judge	Total Score	Total Count	Obsvd Avg	Fair(M) Avg	Infit		Outfit		Corr. PtBis
					MnSq	ZStd	MnSq	ZStd	
Q	417	56	7.45	7.54	.64	-2.0	.65	-2.0	.48
I	413	56	7.38	7.46	.98	.0	1.03	.2	.40
K	358	56	6.39	6.45	.94	-.2	.94	-.2	.36
J	357	56	6.38	6.43	1.12	.6	1.15	.8	.41
F	315	56	5.63	5.64	.88	-.6	.87	-.7	.51
H	308	56	5.50	5.50	.70	-1.8	.71	-1.7	.42
E	291	56	5.20	5.17	.87	-.7	.87	-.7	.53
P	268	56	4.79	4.71	.77	-1.2	.73	-1.5	.49
M	264	56	4.71	4.63	2.05	4.4	2.07	4.5	.27
Mean	332.3	56.0	5.93	5.95	.99	-.2	1.00	-.2	.43
S.D.	54.1	.0	.97	1.03	.40	1.8	.41	1.9	.08

RMSE (Model) 0.13; Adj S.D. 0.94; Separation 6.94; Reliability 0.98;
Fixed (all same) Chi-square: 403.9; d.f.: 8; Significance: 0.00.

Table 1 displays the *judge* measurement report. Columns 'Total Score' and 'Total Count' report the sum of the ratings and the number of ratings that each rater assigned, respectively. Similarly, the fourth column shows the average score each rater assigned ('Obsvd Avg'), and the fifth column shows the average expected rating for each rater ('Fair(M) Avg', the 'fair average' based on the MFRM model).

The rater fit indices, shown in columns six through nine, all indicate that the ratings are consistent with the MFRM model, except for the last judge, Judge M, whose values of Mean Square ('MnSq') for both infit and outfit are well outside of the recommended range 0.6-1.4; the same conclusion is drawn by inspecting the standardized residuals' columns ('ZStd'), where absolute values

larger than 2 only occur with Judge M, signalling unexpected behaviour. This suggests that the most severe judge may have induced a rater effect.

Finally, the SR/ROR correlation, shown in last column 'Corr. PtBis', indicates that the ratings that these raters assigned exhibited a moderately high level of agreement. Judge M, again, reports the lowest value of association.

The rater separation index H can be computed from the output as $(4G + 1)/3 = 9.587$, were G is the separation value 6.94; this suggests that there are about ten statistically distinct strata of rater severity in this sample. As H is larger than the total number of raters, it indicates that the spread of the rater severity measures is considerably greater than the precision of those measures.

The high degree of rater separation reliability (0.98) implies that raters are differentiated in terms of the levels of severity they exercised. There is some evidence here of unwanted variation between raters in their levels of severity. This is also confirmed by the significance of the fixed effect chi-square test reported in the last row of Table 1. The chi-square value of 403.9 with 8 degrees of freedom is statistically significant, meaning that the raters did not all exercise the same level of severity when evaluating ratees. However it is important to emphasize that the rater fixed chi-square test is very sensitive to sample size. As a result, in many applications of MFRM, the rater fixed chi-square statistic may be statistically significant even if the actual variation between raters in the levels of leniency/severity exercised is small (Myford and Wolfe, 2004).

Additional results from Table 2, displaying the *feature* facet measurement report, suggest a possible halo effect. When most of the judges exhibit halo effects, ratings are similar across features. As a result, the features would appear to differ little in terms of their difficulties when the features, indeed, do differ in their true difficulties. The fixed effect chi-square test performed on the *feature* facet is not significant ($p = 0.19$), indicating that there are no significant differences among them in terms of difficulty. Also the separation index H is very small and equal to $(4G + 1)/3 = 0.91$ (where $G = 0.43$), with a reliability of 0.16, implying that raters could not reliably distinguish much among the different aspects. However, the appearance of no difference in feature difficulty does not necessarily imply that the raters exhibited halo. Features can be conceptually distinct but do not differ in difficulty.

Table 2 Feature Measurement Report of the first dataset, U/11C in 2018

Feature	Total Score	Total Count	Obsvd Avg	Fair(M) Avg	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Corr. PtBis
Image	768	126	6.10	6.20	1.07	.5	1.09	.7	.51
Technique	747	126	5.93	6.01	.93	-.5	.92	-.6	.53
Show	739	126	5.87	5.93	1.03	.3	1.05	.4	.54
Coreography	737	126	5.85	5.91	.96	-.3	.95	-.4	.53
Mean	747.8	126.0	5.93	6.01	1.00	.0	1.00	.0	.53
S.D.	12.3	.0	.10	.11	.06	.5	.07	.6	.01

RMSE (Model) 0.09; Adj S.D. 0.04; Separation 0.43; Reliability 0.16;
Fixed (all same) Chi-square: 4.7; d.f.: 3; Significance: 0.19.

5.2.2 Results of Synchro Latin Dance, group U/15C

The first column of the ruler (Figure 4) shows that the logit measure looks more compact than 2018, spanning from -1 to +2 only. The rater severity measures displayed in the second column suggest that the judges do not have very different representations on the logit scale: they all lie within -0.5 and +0.5; however Judges L and U can be identified as the most lenient ones, while Judges G and K as the most severe ones. The feature difficulties appear to be not very different in this case as well; the ordering in the third column suggests that the technique is the most difficult aspect, i.e. it is more difficult to receive high ratings on that feature than it is on the other features appearing higher in the column.

The fourth column includes the performance measures of the athletes. What is surprising is that the order on the logit scale does not reproduce that of the actual final ranking. According to the model, the actual bronze medal should have been awarded as second best, and the real second classified should rank third, jointly with the athlete that got the 6th position. In addition, the bottom ranked athlete, Athlete 11, performed so well as to gain - according to the model - position 5, jointly with the actual fourth and fifth. This suggests that something strange happened during the rating phase, and that the corrections made to the scores according to the regulations have changed the final ranking quite a lot.

Regarding the scale, here the raters extended their scoring to the extreme values: the last column shows clearly that values 1 and 10 were used; however, all the other assigned rates are concentrated in 6, 7 and 8. This may suggest a problem of central tendency. Central tendency may manifest as a lack of variation between ratees in the level of performance demonstrated, as the overuse of the middle categories may harden the distinction among the athletes. This could be partially confirmed by the fact that athlete performances are not entirely distributed along the logit measure. However, results from the fixed effect chi-square test (see Table 3) performed on the *athlete* facet tell that there is a significant difference between at least two athletes; furthermore, the ratee separation index H, here equal to 5.24, confirms that there are about

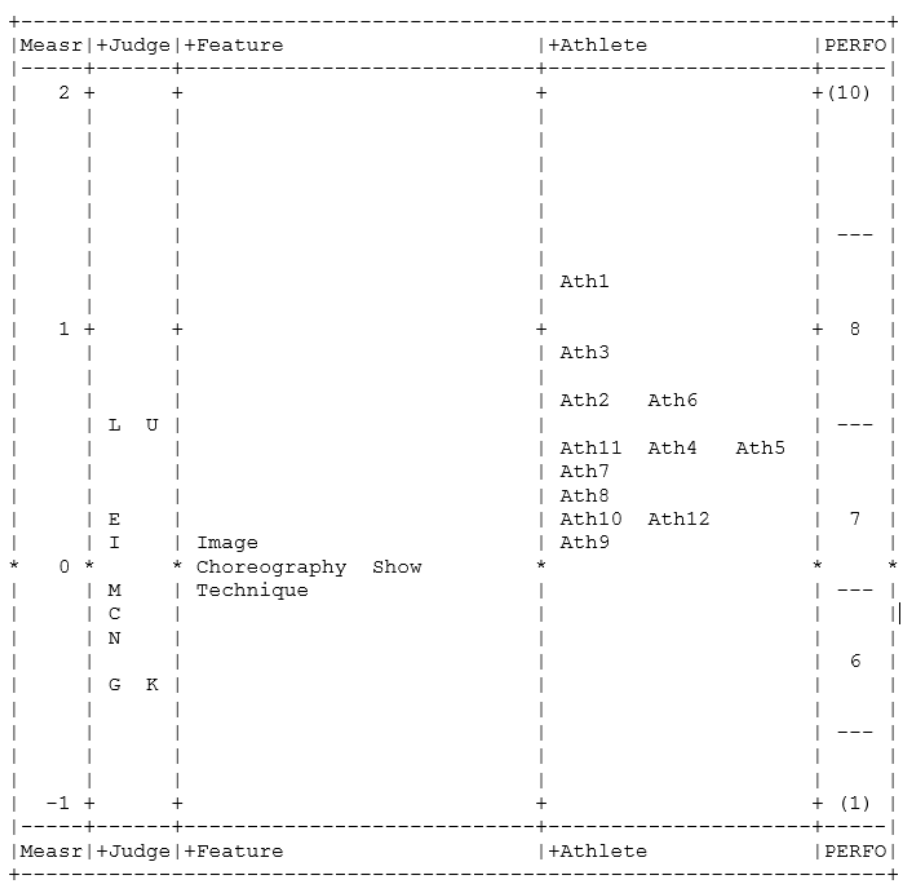

Fig. 4 Variable Map from Facets Analysis of the second dataset [U/15C in 2019], with three facets: athletes, aspects, and judges (raters).

five groups of athletes with homogenous performance. Finally, the separation reliability is equal to 0.79, suggesting that examinees can be differentiated fairly well in terms of their levels of proficiency.

In order to understand better what is underlined by the model, we may now focus on Table 4, that contains the *judge* measurement report. The infit and the outfit mean square measures highlight anomalous values for several raters; in particular Judges U and N have very low values for both fit measures, while Judges E and C very high. The same point can be raised by examining the corresponding standardized residuals that are much larger than 2 in absolute value. It is interesting to notice that their unexpected evaluations are not characterized by a proper tendency of leniency or severity; indeed, as noticed before, judge severity scores on the logit scale are all near (see Figure 4). However, the heterogeneity is confirmed by the correlation values: all of them are very low; the strongest association reaches 0.39 and the average is 0.17, suggesting that there is no wide agreement on the evaluations.

Table 3 Athlete Measurement Report of the second dataset, U/15C in 2019

Athlete	Total Score	Total Count	Obsvd Avg	Fair(M) Avg	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Corr. PtBis
Ath1	298	36	8.28	8.30	.75	-.9	.73	-1.1	.41
Ath3	284	36	7.89	7.92	1.64	2.2	1.65	2.3	.08
Ath6	274	36	7.61	7.65	.42	-3.0	.43	-3.0	.57
Ath2	273	36	7.58	7.62	.78	-.9	.79	-.8	.10
Ath4	266	36	7.39	7.43	1.86	3.0	1.87	3.0	.02
Ath11	266	36	7.39	7.43	1.20	.8	1.21	.9	.08
Ath5	263	36	7.31	7.34	1.46	1.8	1.40	1.6	.18
Ath7	261	36	7.25	7.29	.43	-3.1	.43	-3.1	.46
Ath8	257	36	7.14	7.17	.91	-.3	.89	-.4	.32
Ath12	251	36	6.97	7.00	.83	-.7	.83	-.7	.15
Ath10	248	36	6.89	6.92	.57	-2.2	.55	-2.3	.32
Ath9	239	36	6.64	6.72	1.15	.7	1.15	.7	.28
Mean	265.0	36.0	7.36	7.40	1.00	-.2	.99	-.3	.25
S.D.	15.4	.0	.43	.43	.45	2.0	.45	2.0	.17

RMSE (Model) 0.14; Adj S.D. 0.28; Separation 1.95; Reliability 0.79;
Fixed (all same) Chi-square: 52.2; d.f.: 11; Significance: 0.00.

Table 4 Judge Measurement Report of the second dataset, U/15C in 2019

Judge	Total Score	Total Count	Obsvd Avg	Fair(M) Avg	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Corr. PtBis
U	391	48	8.15	8.16	.35	-4.0	.33	-4.2	.07
L	390	48	8.13	8.14	.78	-1.0	.81	-.8	.15
E	367	48	7.65	7.67	2.45	5.1	2.46	5.1	.22
I	361	48	7.52	7.59	.82	-.8	.79	-1.0	.04
M	346	48	7.21	7.23	.69	-1.7	.69	-1.6	.29
C	344	48	7.17	7.19	1.78	3.3	1.75	3.1	.19
N	337	48	7.02	7.04	.49	-3.2	.49	-3.2	.39
G	322	48	6.71	6.72	.87	-.6	.88	-.5	.01
K	322	48	6.71	6.72	.73	-1.5	.74	-1.4	.16
Mean	353.3	48.0	7.36	7.39	.99	-.5	.99	-.5	.17
S.D.	24.4	.0	.51	.51	.64	2.8	.64	2.8	.11

RMSE (Model) 0.12; Adj S.D. 0.35; Separation 2.85; Reliability 0.89;
Fixed (all same) Chi-square: 76.8; d.f.: 8; Significance: 0.00.

The rater separation index is equal to $(4G + 1)/3 = 4.133$, were G is the separation value 2.85; this suggests that there are about four statistically distinct strata of rater severity in this dataset, far more than would be expected when adopting the standard view with its implied objective of employing raters drawn from a group that is as homogeneous as possible.

The high degree of rater separation reliability (0.89) implies that raters are seriously differentiated in terms of the levels of severity they exercised. This is also confirmed by the significance of the fixed effect chi-square test, that rejected the null hypothesis that all raters exercised the same level of severity when evaluating ratees, after accounting for measurement error.

Also for this dataset, further results from Table 5 point to a probable halo effect. The fixed effect chi-square test performed on the *feature* facet is highly non-significant ($p = 0.52$). Also the separation index H amounts to 1.33, with a reliability of 0.00. Such results suggest that the four aspects under evaluation are homogenous in terms of difficulty or that the judges are not fully able to distinguish between the features.

Table 5 Feature Measurement Report of the second dataset, U/15C in 2019

Feature	Total Score	Total Count	Obsvd Avg	Fair(M) Avg	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Corr. PtBis
Image	810	108	7.50	7.56	.96	- .2	.97	-.2	.27
Coreography	794	108	7.35	7.41	1.07	.5	1.06	.4	.30
Show	792	108	7.33	7.41	.98	-.1	.98	-.1	.27
Technique	784	108	7.26	7.32	.98	-.1	.98	-.1	.27
Mean	795.0	108.0	7.36	7.43	1.00	.0	.99	.0	.28
S.D.	9.4	.0	.09	.09	.04	.3	.04	.3	.01

RMSE (Model) 0.08; Adj S.D. 0.00; Separation 0.00; Reliability 0.00;
Fixed (all same) Chi-square: 2.3; d.f.: 3; Significance: 0.52.

6 Discussion and conclusions

This work highlights how crucial the evaluation process is, particularly in the context of sport performances where there are no right or wrong answers and subjectivity plays a great role. The Many-Facet Rasch model has extended the possibility of objective measurement to examinations which include subjective judgments. If the empirical data cooperate in the construction of a uni-dimensional variable, of the type required to summarize into a single measure an examinee's performance on an examination, then the model is able to provide such a measure on a linear scale with a well-defined standard error. The resulting statistical framework for the analysis of rating data permits summarizing overall rating patterns in terms of group-level main effects for the raters, ratees and features. The contribution of each facet is separated out and examined independently from other facets so as to determine the extent to which various facets are functioning as intended.

Results of MFRM models on the datasets considered here showed clearly that similar competitions can be dealt with in a different way. Judges of the Under 11 competition exhibit a different level of severity but they resort to a coherent ranking, at least at the most extreme positions, where differences should be more evident. MFRM allows identification of the judges which may have adopted anomalous behaviour, introducing a rater effect in the appraisal. In fact, results on the Under 15 competition suggest that a problem in the evaluation occurred: several judges had a non-coherent evaluation of the athletes that could have led to a completely different final ranking.

The comparison of the MFRM results for the two competitions also highlights how the four features are not significantly different in terms of difficulty. Such an outcome can also be a trace of a halo effect, whereby raters could not reliably distinguish between traits. Further in-depth analyses are needed in order to confirm this.

The Many-Facet Rasch models have proved to be an important tool for detecting possible issues in the rating system and have highlighted the importance of an intervention in this field. Some measures have already been enacted, such as the adjustment on the construction of the final ranking by considering the median of individual ranking rather than using the total score. However, some other actions could still be executed. For example, judges could be trained in better use of the whole scale of values and a more effective grasp of differences between the features under evaluation. Future developments can consider more and wider examples so as to confirm what has emerged from this study, and can exploit the behaviour and the consistency of a set of judges in a longitudinal perspective, by combining the outcomes of several competitions. With extensive and solid evidence further actions aimed at improving the rating system for the discipline can be taken.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika 43*(4), 561–573.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge University Press.

Carroll, J. D. and J.-J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika 35*(3), 283–319.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and Evaluating Rater-Mediated Assessments.* Language Testing and Evaluation. Peter Lang Edition.

Engelhard, G. (2002). *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*, Chapter 11, Monitoring raters in performance assessments, pp. 261–287. Routledge.

Farrokhi, F. and R. Esfandiari (2011). A many-facet rasch model to detect halo effect in three types of raters. *Theory & Practice in Language Studies 1*(11), 1531–1540.

Harshman, R. (1970). Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics 16*, 1–84.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing 26*(2), 275–304.

Knoch, U., J. Read, and J. von Randow (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing 12*(1), 26–43.

Linacre, J. (1989). *Many-Facet Rasch Measurement* (I ed.). MESA Press.

Linacre, J. (1994). *Many-Facet Rasch Measurement* (II ed.). MESA Press.

Linacre, J. (2002). Facets, factors, elements and levels. *Rasch Measurement Transactions 16*(2), 880.

Linacre, J. (2009). Local independence and residual covariance: A study of olympic figure skating ratings. *Journal of Applied Measurement 10*(2), 157–169.

Linacre, J. (2013). *Facets computer program for many-facet Rasch measurement, version 3.71. 4*. Beaverton, Oregon: Winsteps.com.

Looney, M. (2004). Evaluating judge performance in sport. *Journal of Applied Measurement 5*(1), 31–47.

Murphy, K. and W. Balzer (1989). Rater errors and rating accuracy. *Journal of Applied Psychology 74*(4), 619.

Myford, C. and E. Wolfe (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement 4*(4), 386–422.

Myford, C. and E. Wolfe (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement 5*(2), 189–227.

Parke, C., S. Lane, and C. A. Stone (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation 12*(3), 239–269.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Roever, C. and T. McNamara (2006). Language testing: The social dimension. *International Journal of Applied Linguistics 16*(2), 242–258.

Saal, F., R. Downey, and M. Lahey (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin 88*(2), 413.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika 31*(3), 279–311.

Wolfe, E. W. and L. Dobria (2008). *Best Practices in Quantitative Methods*, Chapter Applications of the multifaceted Rasch model, pp. 71–85. Sage Thousand Oaks, CA.

Wright, B. D. and M. H. Stone (1979). *Best test design.* Mesa press.