Psychology of Sport and Exercise 18 (2015) 47–52

Contents lists available at ScienceDirect

Psychology of Sport and Exercise

journal homepage: www.elsevier.com/locate/psychsport

A theory-based intervention to prevent calibration effects in serial sport performance evaluations

Frowin Fasold [a, *], Daniel Memmert [a], Christian Unkelbach [b]

[a] Institute of Cognitive and Team/Racket Sport Research, German Sport University Cologne, Germany
[b] Faculty of Human Sciences, Social Cognition Center Cologne, University of Cologne, Germany

ARTICLE INFO

Article history:
Received 9 September 2014
Received in revised form
5 December 2014
Accepted 14 January 2015
Available online 26 January 2015

Keywords:
Calibration
Serial position effects
Range-frequency theory
Interventions

ABSTRACT

Objectives: Serial performance evaluations show calibration effects: Judges avoid extreme categories in the beginning (e.g. *best* or *worst*) because they need to calibrate an internal judgment scale (Unkelbach et al., 2012). Successful calibration is therefore important for fair and unbiased evaluations. A central prerequisite for successful calibration is knowledge about the performance range. The present study tests whether advance knowledge about the range (best and worst) of performances in a series reduces calibration effects.
Design: A 2 × 2 × 2 design was developed with two between subject factors: the knowledge about the performance range (with vs. without) and two different talent tests (specific vs. unspecific). As within subject factor the position of the performances in the series (position 1–10 vs. 11–20) was integrated. The combination of the between subject factors resulted in four experimental conditions.
Method: Handball coaches were randomly assigned to one of the conditions. Afterwards twenty performances were evaluated in a randomized order by the coaches.
Results: Without knowledge about the range, they showed the expected avoidance of extreme categories in the beginning independent of the presented talent test. However, observing the best and worst performance in advance prevented the biases. Range-presentation is therefore a viable theory-based intervention to improve fairness in serial judgments.

In one out of four Olympic disciplines, winning or losing depends on the subjective evaluation of judges or a jury (Stefani, 1998). Further, in talent tests, aptitude tests, or sport examinations, judges[1] evaluate and categorize serial performances based on their subjective impressions. In principle, the subjective character of such evaluations threatens the issue of fairness (Wedell, Parducci, & Roman, 1989), as factors unrelated to the to-be-judged performance might influence evaluations. One of those factors are serial position effects, meaning that performance evaluations are systematically influenced by performances' position in a given competition; one main example of this serial position effect is that performances are evaluated not as good in the beginning as performances in the end (e.g. in gymnastics, Plessner, 1999; or figure-skating, Bruine de Bruin, 2005). A prominent research question is therefore how and when serial position effects arise and how to prevent them.

Calibration in serial evaluations

One possible explanation of serial position effects are *calibration* processes (Unkelbach & Memmert, 2014; Unkelbach, Ostheimer, Fasold, & Memmert, 2012); the calibration explanation assumes that judges must calibrate an internal function that translates observable stimulus input onto available rating systems. As long as this function is not calibrated, judges should avoid extreme categories to avoid consistency violations in the series (Unkelbach et al., 2012; see below). This in turn leads to centering biases in the assessment in the beginning of the judgment series; that is, excellent performances are not judged as good in the beginning compared to the ending and poor performances are not judged as bad in the beginning compared to the ending. Recent research

* Corresponding author. Institute of Cognitive and Team/Racket Sport Research, German Sports University, Am Sportpark Müngersdorf 6, 50933 Cologne, Germany. Tel.: +49 (0) 221 4982 4293.
E-mail address: f.fasold@dshs-koeln.de (F. Fasold).

[1] For simplification we use the term judge ore judges not in the classical sense solely for judging gymnastics or figure skating performances. In this paper judge is used as a synonym for talent scouts, coaches, teachers, or examiners, for every person who has to evaluate performances in series.

http://dx.doi.org/10.1016/j.psychsport.2015.01.001
1469-0292/

(Fasold, Memmert, & Unkelbach, 2012; Unkelbach et al., 2012) has already discussed, that this explanation provides a parsimonious alternative for the mentioned examples of gymnastics (Plessner, 1999) and figure-skating (Bruine de Bruin, 2005), as well for similar biases in other domains (e.g. oral examinations, Colton & Peterson, 1967). Here, we provide a short overview of the calibration explanation, delineate an intervention to prevent serial position effects in evaluations, and test this intervention in a talent scouting test with advanced team-handball coaches. Finally, we discuss the data's implications for the calibration explanation and applications in sport performance evaluations.

The calibration explanation was initially introduced to explain the lack of yellow cards (i.e., an extreme judgment) in the beginning of soccer games (Unkelbach & Memmert, 2008). Further research developed this account into a general explanation of evaluation biases in serial judgments (Fasold et al., 2012, Fasold, Memmert, & Unkelbach, 2013; Unkelbach et al., 2012). As stated above, judges need a transformational function to evaluate performances in serial evaluations. Parducci's range-frequency theory, for example, provides such a function (e.g. Parducci, 1965). Unkelbach and colleagues assumed that the parameters of this function are not fixed but need to develop over the course of a given serial evaluation. They termed this development calibration and as the function has only subjective impressions as input, the only criterion for calibration is the internal consistency of judgments over time (Haubensak, 1992).

An interesting implication of this explanation is that extreme evaluations have a higher likelihood to violate the internal consistency of the function. Imagine someone judging a series of three performances with three categories *good – average – poor*, and judges categorize the first performance as good or poor. However, following performances might be much better or much worse. And consequently judges must use the same category (good or poor) for very different performances, committing a consistency violation. In comparison, the categories average allows at least one further judgment that will for sure not violate judgmental consistency. Thus, extreme categories reduce judgmental degrees of freedom most strongly, leading to higher likelihoods of consistency violations. And as consistency violations are unpleasant (Gawronski & Strack, 2012; Heider, 1958), judges avoid extreme evaluations and judgments until the function is calibrated to the judgment context (see Unkelbach & Memmert, 2014). The calibration explanation thereby locates the cause for serial position effects in a motivational tendency, a need to avoid extreme categories in the beginning. This effect generalizes to any serial evaluation with categorical ratings. Judges evaluate good performances worse in the beginning compared to the end, and poor performances better in the beginning compared to the end. As performances in the beginning, which might be the best or the worst performances of a series, have an a priori lower likelihood to receive extreme ratings, a serious fairness problem arises in serial evaluations.

Improving judgment quality – existing evidence

Apparent judgmental biases in artistic and compositional sports called for more objectivity and transparency in evaluations (e.g. gymnastics, Morgan & Rotthoff, 2014; figure skating, Emerson & Arnold, 2012). For instance, in gymnastics, the mean grades of a judging panel were changed into a complex scoring system with open range of points combining scores for difficulty and execution (gymnastics) and the use of video-based analyses (figure skating) is considered to help judges form more objective evaluations. Experimentally, the employment of fully automated software systems is considered to reduce judgment biases and improve objectivity (Díaz-Pereira, Gómez-Conde, Escolan, & Olivieri, 2014). Despite these efforts, aptitude tests, talent tests, or sport-exams in school settings or university contexts still use subjective serial judgment situations; in these settings, complex algorithms as well as sophisticated technical support are not used due to obvious practical considerations (e.g. the costs of acquiring and maintaining video systems). However, there are possible low-cost and low-effort interventions to ensure that judgments are not unduly influenced by serial position biases.

For example, Unkelbach et al. (2012) tested end-of-sequence assessments; that is, judges assess performances not until they have seen every performance in a series. This procedure prevented the avoidance bias of extreme ratings within in the first performances of an oral examination series. The intervention is based on the assumption that if the complete series is known, judges have a chance to calibrate their transformational function and could assess every single performance without the need to avoid consistency violations. This method is practical and functional for short evaluation series. However, if there are longer judgment series, the final assessment will depend on memory capacity (Engle, 2002). In aptitude or talent tests with a high number of participants, such a strategy is therefore not possible for judges. Additionally, Unkelbach et al. (2012) suggested that end-of-sequence judgments are vulnerable for primacy or recency effects (e.g. Kerstholt & Jackson, 1998; Steiner & Rain, 1989).

The present experiment – a theory-based intervention

Here, we aim to test another strategy that follows from the transformational function suggested by range-frequency-theory (Parducci, 1965, 1968; Parducci & Wedell, 1986). Parducci and colleagues proposed the range principle as one constituent of the judgment function. The range value of a stimulus i in context c is $R_{ic} = (S_i - S_{min})/(S_{max} - S_{min})$, with S_i being the subjective impression of i, S_{min} being the minimal value, and S_{max} being the maximum value in that context. Thus, the best (S_{max}) and the worst performance (S_{min}) of a talent test determine the range of this test, and the range value of each stimulus determined by this difference in the denominator. The range principle explains why the same good performance is judged as *poor* in the context of excellent performances, while it might be judged as *excellent* in the context of poor performances. If the range is known in advance, judges should be able to calibrate these parameters of their judgment function in advance and no centering biases should occur. Judges must not avoid the extreme categories due to possible consistency violations, because the extremes (in our example the best and the worst performances) of the stimulus series are already known.

A problem is how to determine the range of a given context before starting evaluations. To solve this problem, one must assume that the performance levels (e.g. the best and worst performances) are comparable across contexts. For example, in talent tests which are carried out every season, the performance level must be comparable across seasons. That is, given similar tests, the range parameters should be relatively constant over series if the sample of performances is large enough. Given this assumption, an easy way to provide judges with knowledge about the range is the presentation of the range of previous evaluation. With the knowledge of this range, judges should show less centering biases because they already have an important piece of information to calibrate their transformational function.

The following experiment investigates this theory-based intervention with a series of twenty performance evaluations. The experiment thereby simultaneously tests a solution for fairness issues in serial evaluations and tests the calibration explanation of serial position effects. We predict that judges with advanced knowledge of the range of performance do not avoid extreme

ratings at the beginning of a judgment series. Previous research (Fasold, Memmert, & Unkelbach, 2013; Unkelbach & Memmert, 2008) suggests that in a series of twenty evaluations, the transformational function is calibrated after nine to ten awarded judgments.

Based on this previous data, we predict more precisely that first, judges who were shown the range of performance prior to the evaluation process should not avoid extreme ratings, neither during the first ten ratings, nor during further ratings. Therefore, extreme ratings should be randomly distributed according to the random order of extreme performances over the series of evaluations. Second, judges without advanced knowledge about the range should show the *standard* calibration effect; that is avoidance of extreme ratings within the first judgments. Extreme ratings should not be randomly distributed, but should increase from the first ten to the remaining judgments. Based on these two predictions, third, we assume that judges with advanced range-knowledge should use overall more extreme ratings, as judges without advance range-knowledge avoid extreme ratings until they have factually calibrated their judgment scale.

To test our predictions, we used a talent test in team-handball which is conducted every year in a regional team-handball center (located in North Rhine-Westphalia) to scout the best talents of an age-group (under 12 years). This test consists of two standardized evaluations (specific motoric skills and unspecific motoric skills), a free play in a reduced game (four against four), and a free play in the game (six against six with keepers). According to the German Handball Federation, testing specific and unspecific skills generally is part of a talent scouting in German team-handball (Pabst et al., 2013; Schorer et al., 2012). Our study focuses on the standardized procedures because they were conducted at the beginning of the talent evaluation process and all the talents presented their performance in series. We applied our intervention on both standardized tests (specific motoric skills: dribbling course; unspecific motoric skills: gymnastics) to test the robustness of our assumption.

Method

Participants and design

Fifty-seven advanced team-handball coaches (age $M = 25.48$, $SD = 7.80$; 19 female, 39 male) with a mean experience of 4.29 ($SD = 5.67$) years in practice with youth team-handball athletes participated voluntarily in this study; they were naïve with regard to the theoretical background of the experiment. They were randomly assigned to one of four conditions resulting of the orthogonal combination of the factors *test* (specific test vs. unspecific test) and *range-presentation* (with range vs. without range). We used a series of 20 performances and defined the within-participants factor *position* (position 1–10 vs. position 11–20; see Table 1.). Our dependent variable was coaches' *frequency of extreme ratings*. The study was carried out in accordance with the Helsinki Declaration of 1975.

Materials

The above mentioned team-handball center provided videotaped performances of their talent test in the years 2011–2012. We used 20 videos of the specific motoric skill test and 20 videos of the unspecific motoric skill test. Within the specific test, the talents (age 11–12) showed a dribbling course followed by two shots at the goal. Within the unspecific skill test, the talents showed a short gymnastics floor exercise (two rolls forward, two rolls backward, two spin jumps). We already employed both sets of videos successfully in earlier research (Fasold et al., 2012, 2013). The videos were recorded from the perspective of the judges in the real test situation. To provide a range of performances, three independent experts (team-handball coaches, licensed and experienced in youth team-handball more than ten years) selected the two best and the two worst performances of both tests out of a sample of former tests with the same contents (28 videos, 13 dribbling course, 15 gymnastics). The dribbling course videos lasted 27s on average and the gymnastics videos 17s on average. We used the software E-Prime 2.0 to present the videos according to our experimental design. Instructions and videos were presented on a 17″-Dell-Monitor. Participants got an evaluation sheet to provide their judgments. The grading system ranged from 1.0 (best) to 5.0 (worst) with steps of .3 and .7 mapped on the evaluation sheet.

Procedure

The experiment was conducted in a laboratory room. Participants arrived one at a time at the laboratory; the experimenters seated the participants in front of a monitor and after they signed the written informed consent they were randomly assigned to one of the four experimental conditions. First, participants of all four groups (with range/specific test $n = 16$, without range/specific test $n = 15$, with range/unspecific test $n = 15$, without range/unspecific test $n = 13$) got the instruction that 20 videos of a talent-test would be presented to them. Further, they got specific information according to the test content (gymnastics or dribbling course) and that they should assess the shown performance on the delivered judgment sheet immediately after the presentation. After the judgment, they could start the next video by pressing space on the keyboard. The groups without range/specific test and without range/unspecific test started the evaluations immediately. The groups with range/specific test and with range/unspecific test got the information that the two best and the two worst performances from a former test would be presented to them before they could start the test. These videos were not included in the selected test material. After observing these four videos, participants in this group also started the actual evaluations. Importantly, videos were presented in a new randomized order for every participant. After

Table 1
Observed frequencies of extreme and average ratings in the four experimental conditions based position. Significant differences ($p < .05$) to the expected frequency distributions (specific test: extr. = 7.5, aver. = 142.5; unspecific test: extr. = 10.75, aver. = 119.25) are in boldface.

		Specific test (dribbling course)				Unspecific test (gymnastics)			
		With range ($n = 15$)		Without range ($n = 15$)		With range ($n = 13$)		Without range ($n = 13$)	
		Pos. 1-10	Pos. 11–20	Pos. 1–10	Pos. 11–20	Pos. 1–10	Pos. 11–20	Pos. 1–10	Pos. 11–20
Extreme judgments	*n*	11	9	**2**	8	15	14	**5**	9
	%	7.33	6.00	1.33	5.33	11.54	10.77	3.85	6.92
Average judgments	*n*	139	141	**148**	142	115	116	**125**	121
	%	92.67	94.00	98.67	94.67	88.46	89.33	96.15	93.08

the last judgment, participants were thanked and informed about the content of the study. The study lasted 17 min on average.

Results

Prior to analyses, we excluded three datasets from the analyses due to omitted judgments and later corrections. To analyze the frequency of extreme judgments in our design, we coded the best and the worst grades (1.0 and 5.0) as *extreme judgments* (1) and all other awarded grades as *average judgments* (0). We analyzed frequencies in 2×2 tables; to avoid conservative results Barnard's unconditional test was run (Mehrotra, Chan, & Berger, 2003). Given our precise directional predictions, we employed one-tailed tests, given undirected comparisons, we employed two-tailed tests of significance.[2]

Independent of the factors range-presentation and position we found a slight but significant influence of the factor test on the distribution of extreme and average ratings over all judgments. The number of extreme ratings in the specific test (30 of 600 = 5.00%) was significant lower compared with the evaluations of the unspecific test (43 of 520 = 8.27%; $p = .02$). Based on this result and our aim to test the robustness of our predictions with using two different sets of stimuli, we continued analyzing the data separately for the specific and the unspecific test.

As predicted, the data shows that judges with advanced range-knowledge use significantly more extreme ratings, for the specific test (with: 6.67% vs. without: 3.33%; $p = .03$, one-tailed) and as well for the unspecific test (with: 11.15% vs. without: 5.38%; $p = .008$, one-tailed).

Next, we tested whether the extreme ratings in the conditions with range-presentation were equally distributed among the factor position and secondly, whether in the without range-presentation conditions the number of extreme ratings increases from position 1–10 to position 11–20. To test this hypothesis, we calculated the expected frequencies of ratings for the orthogonal combination of the factors range-presentation and position based on the absolute frequencies of extreme and average ratings (specific test: extreme = 30, average = 570; unspecific test: extreme = 43, average = 477). Please note, as performances were always randomized anew, there should be an equal distribution of the ratings among the combination of the factors.

In the specific test, the expected frequency of extreme ratings was 7.5 and for average ratings 142.5. In the unspecific test, the expected frequency of extreme ratings was 10.75 for extreme and 119.25 for average ratings.[3] Next we compared these expected frequencies with the observed use of the extreme and average categories which are presented in Table 1. If our reasoning is correct, judges with advanced range-knowledge should show a random distribution of extreme judgments across the series; that is, they should not significantly deviate from a priori randomness that follows from the random presentation of the performances. Judges without advanced range-knowledge should show systematic deviations from randomness and have systematically less extreme judgments in the beginning compared to the end.

Indeed, in the specific test, participants with advanced range-knowledge did not show any significant deviations from randomness, neither in the first ten trials, $p = .26$, one-tailed, nor in the second ten trials, $p = .34$, one-tailed. Participants without range-knowledge, however, deviated significantly from the expected frequencies within the first ten judgments, $p = .04$, one-tailed. In the second ten trials no differences were obvious, $p = .43$, one-tailed. Undirected pairwise post-hoc comparisons of the observed frequency distributions of the with and the without range condition confirm the results, that without the advanced range-knowledge the number of extreme ratings is significant lower only within the first ten judgments (see Table 1 specific test, position 1–10, $p = .01$; position 11–20, $p = .87$).

In the unspecific test, participants with advanced range-knowledge showed again no deviation from randomness, neither among position 1–10, $p = .26$, one-tailed, nor among position 11–20, $p = .28$, one-tailed. Participants without range-knowledge again deviated from the expected frequencies among position 1–10, but this difference does not reach conventional level of significance, $p = .06$, one-tailed. Thus, we could not support our prediction in this test. Nevertheless, the pairwise post-hoc comparisons of the observed frequencies show the similar pattern of results as in the specific test: the without range-presentation condition leads to a significant lower number of extreme ratings compared with the with-range presentation condition only among the first ten judgments (position 1–10, $p = .02$; position 11–20, $p = .29$).

Summing up, independent of the evaluated performances, if judges saw the range first, their initial judgments did not differ from later judgments, while judges without advance knowledge about the range avoided extreme ratings in their first ten evaluations.

Discussion

Based on the calibration explanation, we have found avoidance of extreme ratings within the beginning of a judgment series, replicating a number of previous experiments (e.g. Fasold et al., 2013). However, as delineated of the theory, providing judges with a priori knowledge about the range of performances prevented this avoidance: with knowledge of performance ranges, judges provided more extreme ratings at the beginning of a judgment series and an equal distribution of extreme ratings over the whole series. That is, ratings in the beginning did not differ significantly from other ratings in the series.

Unexpectedly, we found that judges provided more extreme ratings in the unspecific compared to the specific test. This may indicate that the concrete performance tests are judged differently, which could be due to several factors (e.g. broader performance level in gymnastics).

Independent of this base-rate difference, we found a similar pattern of results for both performance evaluations, showing the general applicability of the calibration explanation. Whereas our analysis could not indicate significant differences in position 11–20 due to the factor position, there was prominent bias in the unspecific test: with range-knowledge participants award noticeably more extreme ratings (five) than participant without range-knowledge (Table 1). We assume that this bias might be due to the randomized order of performance presentation.

Regardless of such noticeable but non-significant biases, the assumed calibration processes predict and explain our results, and thereby, the experiment supports the calibration explanation. Going beyond the theoretical implications, the applied intervention of the previous range-presentation seems to be a time efficient procedure for talent scouts or judges to ameliorate serial position effects in serial judgment situations.

It must be noted, however, that the range of a previous test does not represent the current performance context and the judges calibrate their judgment scale based on the previous context.

[2] For the use of one-tailed tests see Kimmel (1957).

[3] As the used statistical software R allows only testing with natural frequencies, we rounded 10.75 up to 11 and 119.25 down to 119. For the expected frequencies in the specific test we rounded the 7.5 down to 7 and the 142.5 up to 143. We ran the same analysis with the up rounded 7.5 to 8 and the down rounded 142.5 to 142 and the pattern of results was quite similar.

According to our intervention and the definition of the range value of a transformational function introduced in the beginning (e.g. Parducci & Wedell, 1986) with $R_{ic} = (S_i - S_{min})/(S_{max} - S_{min})$, we modified this range value into $R_{icprev} = (S_i - S_{mincprev})/(S_{maxcprev} - S_{mincprev})$; $S_{mincprev}$ being the worst, $S_{maxcprev}$, being the best performance of a previous test (context c_{prev}). If judges start an evaluation with a priori introduced range presentation this range value must be seen as their baseline for an actual performance context. From a practical perspective we can expect that sometimes the average performances will vary from one test to another test. For instance, in one year there can be more talents with outstanding abilities than in the next year. Nevertheless, every year we have outstanding performances and only the number of highly talented athletes varies. We rarely have a year in which the whole sample of talents, even the outstanding talents, differs that much from the previous year that it could not be evaluated in the range of the previous year. Thus, we are sure that judges should keep in mind such possible performance differences, but we think that these differences do not affect our research question.

Although our study represents interesting findings, the results should be considered with some limitations in mind, but we want to give advice on these limits. Only small parts of our data, the extreme ratings, are in focus of our analysis whereas the highest number of ratings (93%) are summarized in the average category. Additional analysis of the data indicated that if we widened our analysis to the next two extreme categories (second best and second worst grade) we would receive 19.55% extreme ratings on which further analysis could be based. Interestingly, in the specific test this does not affect the pattern of results and the predicted effects, whereas in the unspecific test we no longer found the avoidance of extreme categories in the without range-presentation condition. Nevertheless, the use of second best or worst category preserves one degree of freedom either for a following better or worst judgment. Therefore, the idea of calibration as we have introduced could not predict this result. The calibration explanation at the current status only stated that the extreme categories (best and worst) are avoided in the beginning to preserve degrees of freedom for following judgments. As the main aim of talent scouting is to find talents with outstanding abilities, only the extreme performances, the best ones, should be in main focus of evaluation. Only a negligible amount of such talents can be found among the population and therefore it is even more important that judges could evaluate performances on a calibrated judgment scale, as well as in the beginning of a talent test.

Furthermore, we could distinguish between early (first ten) and late (second ten) judgments in our paradigm, but we cannot explain how long the calibration lasted within the different conditions. It is unclear how our manipulations influence the first judgments of the series. Whereas our results support the proposed length of calibration with nine to ten judgments (Fasold et al., 2013; Unkelbach & Memmert, 2008) in the without-range conditions, we cannot exclude that the range-presentation only leads to a shorter calibration phase. Our design with the randomized performance presentation does not allow a determination of the length of calibration. The same considerations could count for the impact of expertise. Previous studies provide evidence that highly experienced examiners (oral examinations, Unkelbach et al., 2012) and highly experienced referees (yellow card decisions, Unkelbach & Memmert, 2008) also show substantial calibration effects. Neither these two studies nor our presented experiment could state, if expertise could influence the length of the calibration process. The calibration explanation cannot predict such proposed expertise effects in a specific context. However, we suspect that experts require a shorter calibration phase, as they are able to better use the pieces of information of every stimulus to develop their transformational function.

Whereas we investigated only the effect of the advanced range of performance presentation on the development of the transformational function of judges, it remains unclear what the effect of any random performance, presented in advance, could be. According to our underlying function $R_{icprev} = (S_i - S_{mincprev})/(S_{maxcprev} - S_{mincprec})$ we could only predict the effect of the range-presentation. No predictions can be made if mean performances are presented. Such predictions would be speculative regarding the actual conjunction of theory (e.g. range-frequency theory) and the calibration explanation. In this experimental study we followed our clear theoretical assumptions, nonetheless, further research should investigate how for instance mean performances could influence calibration processes in advance.

Nevertheless, the results of our study also support the notion that the calibration explanation can be an alternative to explain the transformational function of judges in the subjective serial evaluation of performances or stimuli. Furthermore, the need for consistent judgments is also a constituent of judgment functions and Haubensak (1992) stated that due to the avoidance of consistency violations, the centering biases in the beginning of judgment series arises. Therefore, investigating the consistency of judgments among serial evaluations should be integrated to future research. The paradigm of Fasold et al. (2013), which compared the judgments over one performance on different, systematically varied, positions in a series, could offer a possibility to investigate this additional variable.

Conclusion

Summing it up, our study supports the calibration explanation as an alternative to firstly predict serial position effects in subjective judgments and secondly predict how judges, scouts or examiners could avoid the negative effects of the calibration. Next to end-of-sequences judgments, which are effective to avoid the centering biases within small judgment series (Unkelbach et al., 2012), the range-presentation is now an alternative for longer judgment series based on empirical research. Actual knowledge about the best and the worst performances of previous evaluations seems to be a sufficient method to prevent these judgment biases in the first ten judgments of an actual evaluation. Therefore, this is an easily conducted and time economic method to improve fairness criteria of subjective serial evaluations.

References

Bruine de Bruin, W. (2005). Save the last dance for me. Unwanted serial position effects in jury evaluations. *Acta Psychologica, 118*, 245–260. http://dx.doi.org/10.1016/j.actpsy.2004.08.005.

Colton, T., & Peterson, O. L. (1967). An assay of medical students abilities by oral examination. *Journal of Medical Examination, 42*, 1005–1014.

Díaz-Pereira, M. P., Gómez-Conde, I., Escolan, M., & Olivieri, D. N. (2014). Automatic recognition and scoring of olympic rhythmic gymnastics movements. *Human Movement Sciences, 34*, 63–80. http://dx.doi.org/10.1016/j.humov.2014.01.001.

Emerson, J. W., & Arnold, T. B. (2012). Statistical sleuthing by leveraging human nature: a study of olympic figure skating. *The American Statistician, 65*, 143–148.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23. http://dx.doi.org/10.1111/1467-8721.00160.

Fasold, F., Memmert, D., & Unkelbach, C. (2012). Extreme judgments depend on the expectations of following judgments: a calibration analysis. *Psychology of Sport and Exercise, 13*, 197–200. http://dx.doi.org/10.1016/j.psychsport.2011.11.004.

Fasold, F., Memmert, D., & Unkelbach, C. (2013). Calibration processes in a serial talent test. *Psychology of Sport and Exercise, 14*, 488–492. http://dx.doi.org/10.1016/j.psychsport.2013.02.00.

Gawronski, B., & Strack, F. (2012). *Cognitive consistency: A fundamental principle in social cognition*. New York, NY US: Guilford Press.

Haubensak, G. (1992). The consistency model: a process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 303–309.

Heider, F. (1958). *The psychology of interpersonal relations*. New Jersey, US: Lawrence Erlbaum Associates.

Kerstholt, J. H., & Jackson, J. L. (1998). Judicial decision making: order of evidence presentation and availability of background information. *Applied Cognitive Psychology, 12*, 445–454. http://dx.doi.org/10.1002/(SICI)1099-0720(199810)12:5<445::AID-ACP518>3.0.CO;2-8.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin, 54*, 351–353. http://dx.doi.org/10.1037/h0046737.

Mehrotra, D., Chan, I., & Berger, R. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics, 59*, 441–450.

Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: findings of a difficulty bias. *Economic Inquiry, 52*, 1014–1026. http://dx.doi.org/10.1111/ecin.12074.

Pabst, J., Büsch, D., Petersen, K.-D., Nowak, M., Hamann, F., Armbruster, C., et al. (2013). *Testmanual zur Talentsichtung des DHB 2013* [*Testmanual for talentscouting of the German Handball Federation 2013*]. Leipzig: Eigenverlag.

Parducci, A. (1965). Category judgment. A range-frequency model. *Psychological Review, 72*, 407–418.

Parducci, A. (1968). The relativism of absolute judgment. *Scientific American, 219*, 84–90.

Parducci, A., & Wedell, D. (1986). The category effect with rating scales: number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 496–516.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport and Exercise Psychology, 21*, 131–144.

Schorer, J., Büsch, D., Fischer, L., Pabst, J., Rienhoff, R., Sichelschmidt, P., et al. (2012). Back to the future. A case report of ongoing evaluation of the German handball talent selection and development system. In J. Baker, S. Cobley, & J. Schorer (Eds.), *Talent indentification and development in sport*. New York: Routledge.

Stefani, R. (1998). Predicting outcomes. In J. Benett (Ed.), *Statistics in sport* (pp. 249–275). London: Arnold.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*, 136–142. http://dx.doi.org/10.1037//0021-9010.74.1.136.

Unkelbach, C., & Memmert, D. (2008). Game-management, context-effects and calibration: the case of yellow cards in soccer. *Journal of Sport and Exercise Psychology, 30*, 95–109.

Unkelbach, C., & Memmert, D. (2014). Serial position effects in evaluative judgments. *Current Directions in Psychological Science, 23*, 195–200. http://dx.doi.org/10.1177/0963721414533701.

Unkelbach, C., Ostheimer, V., Fasold, F., & Memmert, D. (2012). A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes, 119*, 103–113. http://dx.doi.org/10.1016/j.obhdp.2012.06.004.

Wedell, D. H., Parducci, A., & Roman, D. (1989). Students perceptions of fair grading: a range-frequency analysis. *The American Journal of Psychology, 102*, 233–248.