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