Organizational Behavior and Human Decision Processes 119 (2012) 103–113

Contents lists available at SciVerse ScienceDirect

Organizational Behavior and Human Decision Processes

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

A calibration explanation of serial position effects in evaluative judgments

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

[a] Department Psychologie, Universität zu Köln, Köln, Germany
[b] Institut für Kognitions- und Sportspielforschung, Deutsche Sporthochschule Köln, Köln, Germany

ARTICLE INFO

Article history:
Received 25 March 2011
Accepted 8 June 2012
Available online 4 July 2012
Accepted by Julie Irwin

Keywords:
Evaluative judgments
Calibration
Serial position effects
Range-frequency theory
Consistency model
Threshold decisions

ABSTRACT

Judges often evaluate stimulus series on dimensions for which no physical scale exists; for example, when judging academic ability in oral examinations. We propose that judges deal with this problem by calibrating an internal judgment scale that maps stimulus input onto available judgment categories. This calibration process implies serial position effects: Judges should initially avoid extreme categories, because using extreme categories reduces judgmental degrees of freedom, thereby increasing the possibility of internal consistency violations. In four experiments, we show that judgments become indeed more extreme later in a series of judgments. Judges evaluated the same good (poor) performances more positive (negative) at the end of a sequence compared to the beginning. Judges' expertise did not prevent the effect, but allowing end-of-sequence judgments reduced serial position effects. We discuss the implications and possible remedies of these calibration effects on judgment extremity.

Introduction

Serial evaluations are a frequent exercise in professional life. Human resources staff must evaluate series of candidates, reviewers series of scientific ideas, referees series of game situations, and college professors series of student performances. These evaluations are done with rating scales, grading systems, or binary decisions ("accept/reject"). However, judges do most of these evaluations on dimensions for which, often by definition, no physical scale exists (e.g., "soft skills"). So how do people manage such serial evaluations?

We propose that judges deal with this problem by calibrating an internal judgment scale that translates stimulus input onto available categories (e.g., hire/reject, funding/no funding, foul/no foul, grades from A to F). This idea is already present in Galilei's insight (cited by Lewin (1931)) that objects have no categorical properties on their own, but only in reference to their context (i.e., heavy vs. light, large vs. small), and is famously featured in Festinger's social comparison theory (Festinger, 1954). As stimuli have no categorical properties per se, but series of evaluative judgments often require the use of categorical rating systems, there is a need for what we call *calibration* (Unkelbach & Memmert, 2008).

Calibration is the development of an internal scale during a judgment series, or, more specifically, the process by which judges learn to use an available category system to judge stimulus input. The present research investigates serial position effects in evaluative judgments that follow from the calibration process: At the beginning of a judgments series, judges do not know the range of the stimuli they will observe (e.g., how good or bad will the candidates in a series be?). This implies that the same stimulus will be judged differently at the beginning compared to the end of a series. We will show such effects in the context of students' oral examinations. We predict that judges evaluate good exams not as positive as they deserve in the beginning compared to the end, and that they evaluate poor exams not as negative as they deserve in the beginning compared to the end. In other words, we predict that judges avoid extreme categories in the beginning of serial evaluations.

Theoretical background

Social comparison research provides one answer how people deal with serial evaluations for which no physical scale exists (Festinger, 1954; Mussweiler, 2003; Suls, Martin, & Wheeler, 2002). The main notion is that people should make judgments comparatively. Though evaluators cannot make absolute judgments, they should be able to tell if one applicant is more qualified than the other, if one grant proposal is better than the other, and if one student has studied her materials more thoroughly than the other. The great number of studies that show social comparison effects testifies to the power of this idea. However, social comparison research is largely silent about the process how internal comparisons are mapped onto judgment dimensions and categories, and how comparison processes influence series of judgments and decisions.

* Corresponding author. Address: Department Psychologie, Immermannstrasse 49-51, 50931 Köln, Germany.
E-mail address: christian.unkelbach@uni-koeln.de (C. Unkelbach).

0749-5978/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.obhdp.2012.06.004

Parducci's range-frequency theory of categorical judgments (1965, 1968) provides a more specific answer to the question how people map observations onto an available category system, for example, observed academic performance to grades from *A* to *F*. The idea is that people follow a range principle, according to which they infer a *psychological range* from the available stimulus input. Within this range, judges set the categories of the available external judgment scale (e.g., *A* to *F*). The range principle makes similar predictions as standard social comparison models. For example, in classes of excellent students, professors will judge good students only as average, and in classes of average students, professors will judge the same good students as excellent, due to the shift in the psychological range.

The frequency principle then assumes that judges use the available categories with equal frequency. If however, input frequencies are not equally distributed, for example, when there are skewed distributions of good, average, and poor students, the range principle and the frequency principles are in conflict, which is resolved by assigning weights to the respective principles.

Range-frequency theory has received much empirical support (e.g., Parducci & Wedell, 1986), but is not often applied outside of abstract stimulus presentations, although there are notable exceptions (e.g., Niedrich, Weathers, Hill, & Bell, 2009; Wedell & Parducci, 1988; Wedell, Parducci, & Roman, 1989). Range-frequency theory has also been criticized for a number of reasons, mainly for the fact that it is unclear how participants make initial judgments, when they had not yet a chance to infer a psychological range.

The consistency-model by Haubensak (1992) proposes a solution for this problem. The model assumes that an internal scale develops quickly during the very first observations; this explains the range effect, when average students are judged better in a class of poor students compared to a class of excellent students. As frequent stimuli (e.g., excellent vs. poor students) have a higher chance to occur early in a series, the scale is centered around these stimuli. The scale is retained and used consistently across a judgment series: "...because absolute judgments are concerned with subjective impressions only, there can be no right or wrong answers. The only criterion judges can use is the internal consistency of their own responses." (Haubensak, 1992, p. 304).

This insight is also apparent in measurement theory, which states that a scale's most important criterion is validity. If people cannot use validity, they might use the second-most important criterion, reliability, which is the core assumption of the consistency model: stimuli of the same intensity on the relevant dimension should be put in the same category across a judgment series. For example, professors have much leeway to define "pass" and "fail" performances, simply because there is no absolute objective criterion. Yet, they should not let one student pass while another student with the same performance level fails. Thus, each judgment takes away judges' degrees of freedoms in using the available category system.

Calibration

Based on these theoretical considerations, we want to introduce the idea of calibration (see also Unkelbach & Memmert, 2008), which allows new predictions for serial judgments that cannot be derived from range-frequency theory, social comparison models, or the consistency model alone. Rather, the calibration idea combines Parducci's (1965) range and frequency principles with Haubensak's (1992) consistency model.

Let us start with the observation that categories in most evaluation systems are not created equal. In most cases, extreme judgments (i.e., using extreme categories) within a rating system have more significant consequences than others. For example, a *D* in an exam means a lower GPA for students, while an *F* means failing the course and repeating it next semester. The consequences of category judgments depend on the context, but extreme categories will often have more important implications than moderate categories.

Extreme judgments thereby reduce more strongly the judgmental degrees of freedom, as they increase the chance of significant consistency violations. Imagine that a professor fails the first student in a series. Maybe by chance, all other students of the day show even worse performances. Conversely, she might give the first student an *A* and all following students show even better performances. If she uses her category system consistently, she must fail all other students as well or give all students an *A*; otherwise she will experience strong consistency violations. Hence, she should avoid the *F* and *A* categories in the beginning, or more general, the extreme categories, to preserve her degrees of freedom, which allows avoiding significant consistency violations.

However, as proposed by range-frequency theory, after a certain number of trials, a psychological range is established and she can use the extreme categories. This process, when judges still try to preserve their judgmental degrees of freedom and try to map the provided stimulus input onto the available category system, is what we call *calibration*: Judges try to establish a psychological range while being consistent with their own previous judgments, which necessitates degrees of freedom preservation. Which categories are avoided depends on the importance of the consequences resulting from the category judgment and the likelihood of consistency violations. In most cases, judges will avoid the extreme categories of an evaluation system (e.g., an *F* in an exam).

Here, we will test the most basic calibration implication: A serial position effect on judgment extremity, because judges avoid extreme categories in the beginning of a stimulus series. Obviously, preserving degrees of freedom in this way is only sensible when we assume the discussed asymmetry in judgment errors that people try to avoid. A professor can grade the first excellent performance as a "*B*+" and a following similar performance as an "*A*", or she can grade the first bad performance with a "*C*−" and a following similar performance with a "fail", but not the other way round (see Spranca, Minsk, & Baron, 1991). The direction of the prediction depends on the kind of error judges try to avoid.

Let us illustrate this with an example when the available judgment system has only two categories, "select" or "reject". If a professor's task is to identify the best student of the day, she will hardly use the "select" category in the beginning (i.e., to preserve her degrees of freedom). Thus, we would predict that she avoids the extreme category and rejects all students in the beginning of a series. Conversely, when her task is to identify the worst student, she will not use the "reject" category in the beginning and select all students in the beginning of the series.

Obviously, judges should follow an end-of-sequence judgment strategy to select the best or the worst in realistic scenarios. However, the example illustrates that which categories are avoided depends on the context of the judgment task. In addition, the example highlights that calibration is most important when judgments and decisions are made along the way, while end-of-sequence judgments should show less serial position effects.

Judgmental uncertainty

Finally, we have to state an important boundary condition for calibration effects, and that is judgmental uncertainty. Again, no

physical scale exists for many judgments. If such a scale exists, for example, the number of correct items in a multiple choice task, calibration processes should have no impact on judgments because there is a strict conversion rule that maps observations (e.g., correct items) onto an available category system (e.g., grades from *A* to *F*). Such conversion rules reduce judgmental uncertainty to zero.

Very extreme observations also reduce uncertainty. This is easily illustrated by the abstract stimuli often used in the perception and psycho-physics literature to show range-frequency effects. If participants have to judge the size of squares appearing on a computer screen as small, average, or large, we would expect calibration effects. However, if the very first square fills the screen or is hardly visible at all, there will be no uncertainty for judges to use the extreme categories "large" or "small", respectively. In an examination situation, an examinee who simply does not answer any questions at all will fail the exam, whether he appears in the beginning or the end of a series.

Another factor that should decrease uncertainty is experience with the judgment situation. One would expect that experienced judges become overall calibrated across many judgment situations – expert judges should be able to infer an overall psychological range in their field of expertise. Yet, different from the case when physical scales exist or the case of highly extreme exemplars, expertise could also lead to stronger calibration effects. Judges might have learned that every judgment situation has its peculiarities and they need their judgmental degrees of freedom. Imagine an expert teacher who starts at a new school – although she might have examined hundreds of students before, she would probably not fail the first student in the first exam, because she simply does not know yet the performance range at the present school. Hence, even experts should show calibration effects when they are in novel circumstances.

If uncertainty exists about how to assign performance observations to an available category system, we expect calibration effects in serial evaluative judgments. This is the case in many applied contexts (e.g., job interviews), academic performances (e.g., oral exams), and sport competitions (e.g., figure skating).

Existing evidence

Serial position effects are indeed often observed in applied contexts. For example, Scheer and Ansorge (1975) showed that gymnastics coaches' evaluations of video-taped performances increase systematically with the performance's position in the judgment series. They attributed this tendency to coaches' expectancies that competing athletes are ranked by ascending ability level.

Similarly, Bruine de Bruin (2005) reported higher performance ratings later in a series for the Eurovision Song Contest and the World and European figure skating contests. Bruine de Bruin and Keren (2003) explained these order effects with unidirectional comparison processes.

In our own research, we showed serial position effects for yellow card decisions in soccer. In soccer, referees should issue yellow cards for rough, dangerous, and unsportsmanlike fouls. Referees must decide for each foul if it falls into the "yellow card" category or not. Following handbook rules, referees should use yellow cards scarcely, similarly to a professor's "fail" category. We predicted and found in experiments and databank analyses that referees indeed avoid this category in the beginning (Memmert, Unkelbach, Rechner, & Ertmer, 2008; Unkelbach & Memmert, 2008).

For all these cases, we can frame the situations as a calibration problem. Judges avoid extreme ratings in the beginning to preserve their judgmental degrees of freedom to achieve consistency across the judgment series. The following experiments will be the first direct experimental test of the serial position effects predicted by the calibration idea.

Preview of the experiments

In the following four experiments, we investigate the predicted initial avoidance of extreme judgments in the context of oral exams. Oral exams are the typical step-by-step judgment situation with sufficient amount of uncertainty leading to the expected serial position effects. The alternative explanations of expectancy effects or unidirectional comparisons do not apply; examiners should have no expectancies about the order of performance in academic oral examinations, and the series contain good as well as poor performances, making unidirectional comparison unlikely as well. We will discuss these alternative accounts in more detail in the General Discussion.

In all experiments, participants judge a sequence of six oral examinations in physical education theory. Two of these examinations were rated as good by professors, two were rated as average, and two were rated as poor. We expect that judges avoid the extreme categories (*A* or *F*) of the grading system in the beginning. Consequently, they should evaluate examinees' good performances at position 1 not as positive as compared to position 5. Conversely, they should evaluate examinees' poor performances at position 1 not as negative as compared to position 5.

This design also allows testing whether grading differences at position 5 are due to social comparisons (or unidirectional feature comparisons) with the preceding stimuli at position 4. Social comparison models would predict that in this concrete case, a stimulus is contrasted away from the preceding stimulus (e.g., Wänke, Bless, & Igou, 2001). That is, exams at position 5 might be simply judged better or worse in contrast to the previous poor or good exam performances.

We believe that such contrast effects must be part of calibration in the beginning, because contrast effects help defining the range of the scale. The range effect in Parducci's (1965) range-frequency theory can be thought of as a generalized contrast effect: Again, the same average student will appear much better in a class of poor students than in a class of excellent students. Successful calibration, on the other hand, should diminish the influence of contrasting exemplars. As participants learn to map their observations to the available categories, the specific influence of the preceding stimulus should decrease. We will test whether the observed effect is due to contrast effects in all four experiments, using the average performances at positions 2 and 6.

Experiment 1 uses video-taped real examinations and introduces the basic calibration effect (i.e., initial avoidance of extreme categories). Experiment 2 replicates the calibration effect with written transcripts of the examinations. Experiment 3 investigates the role of expertise for the effect. Participants were either novices for judging exams in physical education theory (i.e., psychology students), had some expertise (i.e., sport students who already took this examination), or were experts (i.e., professors who teach this subject and regularly administer such exams). Experiment 4 shows that the serial position calibration effect are partially prevented when judgments are made at the end of the sequence, and not step-by-step.

The main manipulation is always whether a good or a poor exam performance is shown at position 1 – and the evaluation of this first exam will be compared across participants with the same exam evaluated at position 5. Thus, the main dependent variables are the grades awarded at the first and fifth position for poor and good performances.

Experiment 1 – Serial position effects in oral exams

Method

Materials

We filmed 15 physical education students during their exams at the Universität Heidelberg; all previously agreed to being filmed and we ensured that they felt no pressure to participate in this procedure. Using video-editing software, we also ensured that the examinees could not be recognized in the actual experiment.

Each exam lasted around 45 min and the grading ranked from "good" to "fail". From the initial sample, we selected two good exams, two average exams, and two poor exams, based on the assigned grades. We cut these six videos down to 5-min clips and pixelated examinees' faces to protect their identity. From these six videos we compiled four sequences: In the good performance condition, participants judged good exams at positions 1 and 5, while poor exam performances were presented at positions 3 and 4. In the poor performance condition, participants judged poor exams at position 1 and 5, while the good exam performances were presented at positions 3 and 4. Orthogonally, we varied which of the two good/bad exams were used at positions 1 and 5. In all four conditions, the average exams appeared at positions 2 and 6.

Participants and design

Eighty-five psychology students participated either for payment or partial course credit (74 women, 11 men; mean age = 22.64). The design included performance level (good vs. poor performances at position 1) as a between-participants variable and serial position (position 1 vs. position 5) as a within-participants variable. Thus, the comparison of the same exam at position 1 and 5 is done between-participants. As a control factor, we systematically varied which of the two exams of each performance level was used at positions 1 and 5, resulting in the four sequences described above. Participants were randomly assigned to one of the resulting four experimental conditions.

Procedure

After arriving in the laboratory, experimenters seated participants in individual cubicles and started a Visual Basic computer program. This program presented instructions, played the exam clips, and recorded participants' grades for each exam. Instructions told participants to imagine being physical education theory examiners with the task to grade a sequence of verbal exams. The length of the series was not specified. The available grades ranked from 1.0 (*A*) to 4.3 (*F*), using the German university grading system, with lower numbers indicating better grades. For each of the six exams, participants first watched the examination video and then immediately graded the performance by entering the grade into a text field. After participants graded all six exams, the experimenter thanked and thoroughly debriefed them about the hypothesized calibration effect. Experimental sessions lasted about 45 min and up to four people participated per session.

Results

First, we checked how participants overall graded the exams; for the good and poor performances, participants graded the 5-min versions in accordance with the original examiners; they graded good performances better ($M = 1.70$, $SD = 0.38$) than poor performances ($M = 3.57$, $SD = 0.45$). This clear difference in grades shows that participants clearly distinguished between performance levels. Unexpectedly, they graded the originally average performances as good as and even better than the originally good exam performances ($M = 1.47$, $SD = 0.27$). However, this poses no problem for the proposed calibration effect, as we only predicted that participants avoid extreme categories at position 1 compared to position 5. To show this effect, we computed a difference score between each exam's mean rating and participants' individual ratings; in other words, we centered the exam grades around zero, which makes the pattern of means very easy to interpret[1]: This difference score is positive when exam grades are better than their mean evaluation and negative when exam grades become worse than their mean evaluation. Hence, positive scores for good exams and negative scores for poor exams indicate more extreme judgments.

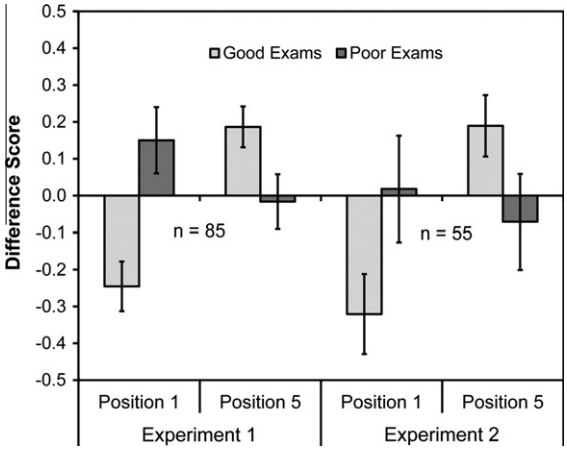

Fig. 1. Deviations from exams' mean grade as a function of performance level (good vs. poor) and position (first vs. fifth) in Experiments 1 and 2. Positive values for good performances and negative values for poor performances indicate more extreme grades. Error bars represent standard errors of the means.

Serial position effects on judgment extremity

The resulting means are presented in the left part of Fig. 1, for good and poor exam performances at positions 1 and 5. As the figure shows, participants rated good performances worse at position 1 ($M = -0.25$, $SD = 0.48$) compared to position 5 ($M = 0.19$, $SD = 0.37$), while they rated poor performances better at position 1 ($M = 0.15$, $SD = 0.58$) compared to position 5 ($M = -0.02$, $SD = 0.48$). In other words, participants' grades for the same exams became more extreme as a function of position in the sequence.

We analyzed these data using a 2 (performance level: good vs. poor) × 2 (position: first vs. fifth) × 2 (control factor: which of the two good/poor performances is presented at position 1) ANOVA, with repeated measures on the second factor. This analysis shows the predicted interaction of performance (good vs. poor) and position (first vs. fifth), $F(1,81) = 29.45$, $p < .001$, $d = 1.21$.[2] This interaction is clearly visible in Fig. 1. In addition, there was an absolute position effect; exams at position 1 were graded slightly worse ($M = -0.05$, $SD = 0.55$) than exams at position 5 ($M = 0.10$, $SD = 0.43$), $F(1,83) = 5.71$, $p < .05$, $d = 0.52$. This effect is due to the stronger serial position effect for good performances. The performance main effect (as the grades are centered on zero) and all other main and interaction effects were negligible, all Fs < 2.0, *ns*. Thus, following the calibration prediction, participants rated the performances less extreme in the beginning compared to the end.

[1] The ANOVA effects remain exactly the same whether we use raw grade scores instead of the centered ratings, as the difference score a linear transformation. Centering the grades makes the pattern of means easier to interpret (see Rosnow & Rosenthal, 1989).

[2] For all significant effects involving one numerator degree of freedom, we provide Cohen's d as an effect size indicator, corrected for sampling error using the formula suggested by Thompson (2006).

Contrast effects

The design also allows testing contrast effects due to social comparison as an explanation for the more extreme judgments toward the end. At position 5, participants judged good/poor performances after having seen poor/good performances, respectively. Thus, we analyzed the grades of the average exams, which were always presented at positions 2 and 6. The contrast logic is as follows: Average exams should appear better following poor exams than following good exams. This reasoning is valid although the average exams were overall rated as good. In the beginning of the series (i.e., position 2), when participants are not yet calibrated, we expect such contrast effects, while they should decrease towards the end (i.e., position 6). The variable of interest is the grading difference between average exams following good exams and average exams following poor exams.

Indeed, at position 2, the difference is positive ($M_{Diff} = 0.22$, $SD = 0.40$) and significantly different from zero, $t(83) = 2.53$, $p < .05$, $d = 0.56$, indicating a contrast effect. Participants graded average exams better following poor exams compared to the exact same exams following good exams. This contrast effect is completely gone and even slightly reversed at the end of the series ($M_{Diff} = -0.05$, $SD = 0.31$), $t(83) = -0.75$, *ns*. Note that at position 6, participants rated the preceding good/poor exam even more positive/negative compared to position 2. Yet, there is no contrast effect for the average exam at position 6. Thus, at position 6, it did not matter anymore whether average exams were preceded by good or poor exams.

Discussion

We found the predicted serial position effect in sequences of oral exams. Participants judged good exams in the beginning not as positive as when the same good exams were presented at the end of the sequence. Conversely, poor exams in the beginning were judged not as negative as the same poor exams at the end of the sequence. This indicates that participants avoided the grading scale's extreme categories in the beginning of the series.

At position 1, participants could not know yet that the observed performance is indeed among the best (or worst) within the series. Thus, although they clearly discriminated between good and poor performances, their ratings were not as extreme in the beginning compared to the end. We believe this uncertainty in the beginning leads to more average judgments because judges try to avoid consistency violations, and possible consistency violations following extreme judgments are more severe and more likely than violations following moderate judgments.

Fig. 1 clearly shows the increased extremity due to serial position. The figure also shows that the effect is asymmetrical. Good exam performances were rated much better later in the series, while poor performances were graded only slightly worse. One explanation is that participants were reluctant to assign extremely negative grades, even at the end of the series. Avoidance of extreme negative judgments is often observed when participants have to provide evaluative judgments in ability domains (Martijn, Spears, Van der Pligt, & Jakobs, 1992; Skowronski & Carlston, 1987). In addition, participants might have been reluctant to award fellow students a 4.3 grade ("fail").

One might argue that the observed interaction is due to localized contrasts. Yet, the analysis of average exams makes this alternative explanation unlikely. In the beginning (i.e., position 2), we do find contrast effects; participants graded moderate performances better following poor exams compared to when the same performance followed good exams. Yet, this contrast effect virtually disappeared at the end of the stimulus series, although we observed more extreme ratings of the good and poor exam performances.

To be sure, we do not deny the existence of contrast effects; we believe they play a major role in establishing the psychological range of the stimulus series. However, while many studies have focused on and found assimilation and contrast for two to-be-judged stimuli (e.g., Herr, Sherman, & Fazio, 1983; Mussweiler, 2003), we conceive contrasts between stimuli as the tool by which a psychological range is established and calibration is enabled. After successful calibration, we expect that the informational impact of specific preceding negative or positive information is diminished.

Experiment 2 – Serial position effects in transcribed exams

The exam videos used in Experiment 1 have face-validity, as they emulate real oral examinations. However, many judgment tasks involve written materials. To investigate if we can replicate and extend the serial position effects to written materials, we used exam transcripts as basis for grading.

Method

Materials

As Experiment 1's student evaluation of the average performance differed from the professors' evaluations, we re-cut the average exams from the original sample of 15 exams into three different versions each, resulting in six average exams. Ten additional students graded these six exams. The grades were consistent across judges (Cronbach's α between .75 and .87). We selected two exams with the most average mean grade (i.e., 2.3) for the main experiment. We transcribed these two average performances, together with the good and poor performances from Experiment 1. The transcripts contained everything the examinees said except fillers like "ehm" and breaks. They were about two pages long with a 12pt Arial font and single-spaced formatting. Different from Experiments 1, materials were presented in paper and pencil format.

Participants and design

The design was the same as in the previous experiment; it included performance level (good vs. poor performances at position 1) as a between-participants variable and serial position (position 1 vs. position 5) as a within-participants variable. We again systematically varied between-participants which of the two exams of each performance level was used at positions 1 and 5, resulting in four experimental conditions. Fifty-five Universität Heidelberg students from various faculties participated for payment of 3 Euros (47 women, 8 men; mean age = 22.24) and were randomly assigned to one of the four experimental conditions.

Procedure

Procedures were similar to the previous experiment, but instructions, exam transcripts, and grading sheets were provided in paper form. To ensure the sequential evaluation in the proper order, participants had two filing baskets in front of them. Exams were in the left one and participants had to take one, read it, and grade it in an answer-booklet, given a possible range of 1.0 (best grade) to 4.3 (fail). The answer-booklet had one page for each exam. After grading them, participants put the exam into the right filing basket and were instructed not to check it again. After completing the six exams, experimenters thanked, paid, and thoroughly debriefed participants. Sessions lasted between 30 and 35 min, depending on individual speed.

Results

Participants' grades were now in agreement with the professors' grades: they graded good performances ($M = 1.85$, $SD = 0.44$) better than average performances ($M = 2.33$, $SD = 0.45$) and poor performances ($M = 3.11$, $SD = 0.45$). Similar to Experiment 1, we computed difference scores between an exam's mean evaluation and participants' ratings. Positive values again show improved grades (i.e., more extreme evaluations for good performances) and negative values show worsened grades (i.e., more extreme evaluations for poor performances).

Serial position effects on judgment extremity

The right part of Fig. 1 presents these centered ratings for good and poor performances at positions 1 and 5. As in Experiment 1, participants graded the same good performances better (i.e., more extreme) at position 5 ($M = 0.19$, $SD = 0.43$) compared to position 1 ($M = -0.32$, $SD = 0.56$), while they graded the same poor performances slightly worse at position 5 ($M = -0.07$, $SD = 0.69$) compared to position 1 ($M = 0.02$, $SD = 0.76$), resulting in the predicted performance level (good vs. poor) by position (first vs. fifth) interaction in the respective ANOVA, $F(1,53) = 6.22$, $p < .05$, $d = 0.68$. Different from the first experiment, the position main effect was not significant on a standard alpha level, $F(1,53) = 3.07$, $p = .085$. All other effects were negligible, $Fs < 1.0$, *ns*.

Contrast effects

When we computed the differences for average exams at positions 2 and 6, following good and poor exams, we observed no contrast effects. There was neither a significant difference as a function of the preceding exams at position 2 ($M_{Diff} = 0.013$, $SD = 0.605$), $t(53) = 0.08$, *ns*., nor at position 6, ($M_{Diff} = 0.179$, $SD = 0.603$), $t(53) = 1.10$, *ns*.

Discussion

Experiment 2 showed the predicted serial position effect on judgment extremity for transcripts of the exams. Replicating Experiment 1, good exam performances were graded better towards the end compared to the beginning. Conversely, poor exam performances were graded slightly worse in the end compared to the beginning. This extends the predicted calibration effect to written materials, although we have to acknowledge that we prevented judges from going back and changing initial grades – a possible strategy when written materials are available.

The serial position effect on judgment extremity is somewhat weaker in Experiment 2 compared to the first experiment, as indicated by the effect sizes ($d = 0.68$ vs. $d = 1.21$). There are two possible causes: First, written materials might provide a better basis for participants' judgments, thereby reducing the uncertainty about the correct judgment category (i.e., the appropriate grade). As uncertainty about the mapping of the category system to the observations is a pre-requisite for calibration effects to emerge, written materials might have helped participants to reduce uncertainty. A second possibility is that the transcripts did no longer convey performance levels as extreme as the videos; this possibility is supported by the range of the mean ratings. Experiment 1's mean grades ranged from 1.70 to 3.57, while Experiment 2's mean grades ranged only from 1.85 to 3.11. Thus, the transcribed exam performances might have been less extreme to begin with. As the predicted serial position effects are mostly visible for decisions involving extreme categories, the overall effect might be reduced here, because fewer participants used extreme categories at both positions 1 and 5.

In addition, we are again confident that the more extreme ratings toward the end are not due to contrast effects. The transcribed average exams were not systematically influenced by the preceding exams. Thus, social comparison seemed to play no role here, which might also be a power problem created by the smaller sample size compared to Experiment 1. Yet, we still observed systematic serial position effects on judgment extremity.

Experiment 3 – Does expertise help?

We already introduced the idea that experience with the judgment situation might be a crucial factor to reduce the observed serial position effects. Expert judges should (a) experience less uncertainty, and (b) might already have developed a conversion rule that maps observed performances onto available categories. On the other hand, expert judges might feel the need to preserve their judgmental degrees of freedom more strongly than novice judges, and thus, show even stronger serial position effects.

To investigate whether expertise does prevent or promote calibration effects, we recruited real examiners to evaluate the videotaped exams. We recruited 16 examiners who regularly administer oral exams in physical education theory. In addition, we also recruited sport students who already had passed an exam themselves in physical education theory. As a comparison group, we again recruited psychology students. The latter can be seen as novice judges, the sport students as semi-experts, while the examiners are true experts for the pertinent judgment situation.

Method

Participants, design, materials, and procedure

We were able to recruit 16 examiners (5 women, 11 men; mean age = 33.44) from the Deutsche Sporthochschule Köln (German Sport University Cologne), 48 sport students from the Deutsche Sporthochschule Köln (13 women, 35 men; mean age = 25.77), and 59 psychology students from the Universität Heidelberg (41 women, 18 men; mean age = 22.83). We used the same six pretested videos we created for the transcripts in Experiment 2. The design was identical to Experiment 1, with the additional factor expertise (novices vs. semi-experts vs. experts). Participants were randomly assigned to one of the four judgment sequences. Procedures were also similar to Experiment 1, with the exception that psychology and sport students evaluated the exams in groups with up to four participants per session. Examiners completed the task individually in their offices on a portable computer provided by the experimenter; the experimenter was blind to the examiner's respective experimental condition.

Upon completing their evaluation of the exams, all participants were thoroughly debriefed about the hypothesized calibration effects, thanked, and paid. Students received 5 Euro for their participation, while examiners were compensated with 50 Euro.

Results

The overall grades reflected student's performance in the examination. Good exams were graded better ($M = 1.86$, $SD = 0.38$) than average exams ($M = 2.35$, $SD = 0.51$) and poor exams ($M = 3.75$, $SD = 0.44$), with lower numbers indicating better grades. The overall grades also showed a linear trend for expertise: Experts graded exams worse ($M = 2.80$, $SD = 0.34$) than semi-experts ($M = 2.69$, $SD = 0.35$) and novices ($M = 2.58$, $SD = 0.29$), $F(1,120) = 6.03$, $p < .05$, $d = 0.45$.

Serial position effects on judgment extremity

We again centered the grades on zero as such that positive values indicate better grades. Fig. 2 presents the relevant means for the three levels of expertise. Fig. 2 shows the predicted position effect at all levels of expertise, with experts visually showing the strongest serial position effects. We analyzed these means using a 2 (performance level: good vs. poor) × 2 (position: first vs. fifth) × 2 (control factor) × 3 (expertise: novices vs. semi-experts vs. experts) ANOVA, with repeated measures on the second factor. This analysis confirmed the interaction of performance and position visible in Fig. 2: As in Experiments 1 and 2, participants rated good performances worse at position 1 ($M = -0.42$, $SD = 0.55$) compared to position 5 ($M = 0.24$, $SD = 0.37$), while they rated poor performances better at position 1 ($M = 0.18$, $SD = 0.46$) compared to position 5 ($M = 0.06$, $SD = 0.57$), $F(1,117) = 37.81$, $p < .001$, $d = 1.14$. There was no interaction with expertise level, $F(2,117) = 0.67$, *ns*. The theoretical considerations predict that experts should differ from the novices: As the overall expertise interaction has two degrees of freedom, we coded a contrast comparing the interaction for semi-experts and experts against the novices. This contrast was also not significant, $F(1,117) < 1$, *ns*.[3]

In addition, when we analyzed the interaction's strength separately for each expertise group, all groups showed significant interactions of performances and serial position, $F(1,57) = 44.68$, $F(1,46) = 9.64$, and $F(1,14) = 5.32$, $ps < .05$, for novices (psychology students), semi-experts (sport students), and experts (examiners), respectively. All expertise groups showed significantly the expected serial position effects, but the strength of this effect did not significantly differ between levels of expertise.

Besides this main result, the ANOVA showed two more effects. Overall, there was again an absolute position effect; exams at position 1 were graded worse ($M = -0.12$, $SD = 0.59$) than exams at position 5 ($M = 0.15$, $SD = 0.49$), $F(1,117) = 11.57$, $p < .001$. This position effect is due to the stronger position effect for good performances compared to poor performances. As Fig. 2 shows, this asymmetry is most pronounced for sport students, who showed only slightly worse ratings for poor performances at position 5 compared to position 1. Similarly, there was an absolute effect for performance level: Grades were overall worse when sequences started with good performances ($M = -0.09$, $SD = 0.35$) compared to when they started with poor performances ($M = 0.12$, $SD = 0.42$), $F(1,117) = 5.22$, $p < .05$., $d = 0.41$.

Contrast effects

Similar to the previous experiments, we tested whether average exams are judged better or worse following good and poor exams. Across all expertise levels, we found clear contrast effects at position 2: Participants judged average exams better following poor exams compared to good exams ($M_{Diff} = 0.71$, $SD = 0.57$), $t(121) = 6.88$, $p < .001$, $d = 1.25$. These contrast effects were completely gone at position 6 ($M_{Diff} = 0.08$, $SD = 0.62$), $t(121) = 0.70$, *ns*. This pattern of significant contrast effects at position 2 and no effect at position 6 was also stable within each expertise level.

Discussion

We replicated the serial position effects from the previous two experiments. Participants did not judge good and poor performances as extreme in the beginning as they did at the end. Importantly, examiners who regularly administer verbal exams in physical education theory showed the same effects as sport students who already took this verbal exam and psychology students. Overall, given the present sample sizes, we could not find any significant differences in the serial position effects' strength across the levels of expertise.

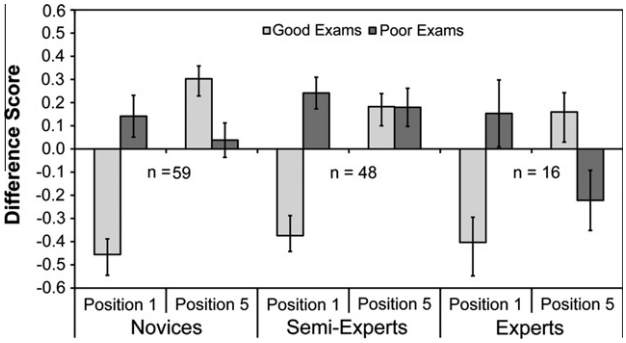

Fig. 2. Deviations from exams' mean grades as a function of performance level (good vs. poor), position (first vs. fifth) and expertise (psychology students as novices, sport students as semi-experts, and examiners as experts) in Experiment 3. Positive values for good performances and negative values for poor performances indicate more extreme grades. Error bars represent standard errors of the means.

Fig. 2 also shows that examiners made the harshest judgments for poor performances later in the series, while sport students only showed a slight decrease in the grades. This pattern helps explaining the observed but not predicted asymmetrical pattern for good and poor performances (i.e., stronger effects for good compared to poor performances). In particular sport students who already took the exam might be reluctant to award other students a "fail", while real examiners have less qualms about letting students fail an exam.

In addition, the proposed calibration explanation for the observed serial position effects is again supported by the differential contrast effects for average exams at positions 2 and 6, which we already found in Experiment 1, but not in Experiment 2. In the beginning, preceding good or poor performance strongly influenced judgments. However, towards the end, participants were no longer influenced by the immediately preceding exams.

Experiment 3 thus shows that expertise does not help per se; but we should not dismiss expertise too fast, but rather ask: expertise for what? For sure, it should be possible to generalize internal conversion rules for observed input to categorical judgments – yet, the context must be comparable if not identical. The examiners in the present experiment had no background knowledge about the examinees, and thus might have been extra cautious with extreme judgments in the beginning. They might have been well aware that all following performances might be better or worse, and thus, they preserved their judgmental degrees of freedom. However, if examiners judge students from the same course within the same contexts, they should be able to preserve their scale for days, weeks, or years. Yet, if context factors (courses taught, study time allowed, etc.) that might influence performance levels change, or the context is unknown, experts might especially try to preserve their judgmental degrees of freedom. This interpretation also fits with the calibration effects observed by Memmert and colleagues (2008) in a databank analysis that concerned only expert referees.

So far, we have found the predicted serial position effect repeatedly and consistently. This consistency might raise doubts about the proposed calibration explanation. There might be some factors inherent in the materials that create these position effects which are unrelated to the proposed psychological explanation proper. The last experiment will therefore show that it is indeed possible to largely reduce the serial position effect with end-of-sequence judgments.

[3] From a visual inspection, Fig. 2 might suggest that the semi-experts differed from the other groups. We also coded an exploratory contrast comparing the semi-experts against the novices and the experts; this contrast also yielded non-significant results, $F(1,117) = 1.27$, *ns*.

Experiment 4 – End-of-sequence judgments

Our theoretical reasoning is as follows: Until judges calibrate an internal scale, they avoid extreme categories in the beginning to preserve their judgmental degrees of freedom, because they do not know how good or bad performances will get in the sequence. In addition, they did not even know the length of the sequences in the previous present experiments. Thus, if judges observe all performances first and then make a series of judgments, we would expect no serial position effects, as judges can calibrate their scale while observing the performances. Once they have seen all performances (and also know the length of the series), there is no further need to preserve judgmental degrees of freedom, and they can use extreme categories in the beginning. Hence, based on the calibration idea, end-of-sequence judgments represent an intervention that should prevent serial position effects.

Method

Participants, design, materials, and procedure

Fifty-three Universität Heidelberg students from various faculties participated in this experiment either for payment or partial course credit (40 women, 13 men; mean age = 22.08). The design was similar to the previous experiments and participants were randomly assigned to one of the four experimental conditions (good vs. poor exam at position 1 × control factor). We used Experiment 3's versions of the oral exams. Procedures were also highly similar to the previous studies, with the exception that participants were informed at the beginning to observe a number of oral exams and judge them afterwards. Thus, instead of judging each exam immediately (i.e., step-by-step), we implemented end-of-sequence judgments. After participants made their six judgments, they were thoroughly debriefed, thanked, and paid.

Results

Overall, participants graded the exams comparable to previous experiments. Good exams were graded best ($M = 1.80$, $SD = 0.48$), average exams in the middle ($M = 2.07$, $SD = 0.50$), and poor exams worst ($M = 3.52$, $SD = 0.55$). Thus, the end-of-sequence judgments did not interfere with participants overall ability to classify the performances.

Serial position effects on judgment extremity and contrast effects

We again centered the grades on zero and analyzed them with the same mixed ANOVA as in Experiments 1 and 2. Different from all previous experiments, this ANOVA showed no significant effects at all. Especially the interaction of performance level and serial position was no longer significant, $F(1,51) = 0.83$, *ns*., as well as all other effects, all Fs < 1. The basic pattern is still visible in the means, though: good performances improve from position 1 ($M = -0.04$, $SD = 0.47$) to position 5 ($M = 0.03$, $SD = 0.72$), while poor performances are rated worse at position 5 ($M = -0.07$, $SD = 0.53$) compared to position 1 ($M = 0.03$, $SD = 0.51$). Yet, these effects are marginal in comparison to Experiments 1 to 3 and far from significant (see above). Given that we found the serial position effect in all experiments with samples as small as $n = 16$, this null finding is unlikely to be a power problem.

In addition, there were no significant contrast effects for average exams following good and poor performances, neither at position 2 nor position 6, $t(51) = 1.33$ and $t(51) = 1.18$, *ns*, respectively.

Discussion

The theoretically derived intervention was successful. When participants observe all performances of a judgment sequence before they make their judgments, the previously observed serial position effect vanishes. We believe observing the full sequence reduces the need to preserve judgmental degrees of freedom and allows judges to use the extreme categories for the full sequence. In other words, seeing the full sequence enables judges to calibrate their internal scale, the rule that maps observed input onto an available category system. The lack of contrast effects for average exams at position 2 and at position 6 further supports this interpretation.

However, end-of-sequence judgements are no universal remedy for serial position effects for at least two reasons. First, they suffer from other unwanted influences, such as primacy and recency effects (e.g., Kerstholt & Jackson, 1998; Steiner & Rain, 1989). And especially for longer sequence, end-of-sequence judgments place a high burden on memory. These effects might be avoided by using appropriate annotation and adjustment systems. The second reason is more problematic: Many situations simply do not allow for end-of-sequence judgments. Oral exams in academic settings, performance evaluations in sport settings, or sentencing in judicial settings are not done after full sequences are observed, but immediately after the exam, the performance, or when the trial is over. Although we would expect end-of-sequence judgments to ameliorate the observed serial position effects, they are often not possible due to the strict procedural protocols of many judgment situations.

General discussion

The present data show that judgments become more extreme as a function of serial positions. In particular, good students were judged better and poor students were judged worse in the end compared to the beginning. We predicted and explained this serial position effect by a process that we termed calibration; this calibration idea is a mixture of Parducci's (1965) range principle and Haubensak's (1992) consistency model. As judges try to be consistent, and extreme judgments have a higher probability to violate the internal consistency of a scale, they avoid extreme categories in the beginning until the psychological conversion rule that maps observed input onto available categories is established. Consequently, judgments of the same performance become relatively more extreme toward the end compared to the beginning of a sequence.

This extremity effect is not predicted by range-frequency theory (Parducci, 1965) or the consistency model (Haubensak, 1992). The former is largely silent about how initial judgments are made and more tailored to explain effects in longer judgment sequences[4] (e.g., 50 to-be-judged stimuli in each condition

[4] Nevertheless, range-frequency theory allows computing predicted values for the presented series. According to Wedell and Parducci (1988), a stimulus *i*'s *frequency* value is a function of its rank in context *c* and the total number of stimuli in context *c*, N_c: $F_{ic} = (r_{ic} - 1)/(N_c - 1)$. The *range* value is determined by the subjective evaluation of the stimulus and minimal and maximal stimulus values in that context: $R_{ic} = (S_i - S_{min})/(S_{max} - S_{min})$. The judgment is determined by a weighted average of the frequency and range value, $J_{ic} = wR_{ic} + (1 - w)F_{ic}$, with w usually set at .5. If we assign good, average and poor performances subjective values of 3, 2 and 1, respectively, these formulas indeed predict improved ratings for good performances later in the series, but no decrement for poor performances. Rather, they predict slight improvements or same level judgments for poor performances, depending on how tied ranks are handled. As one reviewer suggested, one might set a starting point of 0.5 for the first judgment's *frequency* value instead of 0, which seems appropriate for short sequences. This change indeed leads to predicted decrements for poor performances, but not to improvements for good performances. Thus, range-frequency theory does not predict the observed pattern in the present experiments.

of Parducci & Wedell, 1986, Experiment 1). The consistency model posits that due to the need for consistency, initial judgments determine the use of the available categories. One might describe the consistency model and range-frequency theory as purely cognitive models to explain category rating effects. The proposed calibration idea and the resulting need to preserve judgmental degrees of freedom highlights a motivational aspect – judges avoid extreme categories because they should lead with greater probability to consistency violations. This motivational aspect is illustrated by a recent study by Fasold et al. (2012). Gymnastic coaches judged only a single performance; half of them believed they would judge only this one performance, while the other believed they would judge eight successive performances. The latter group significantly avoided the extreme categories and judged the performance as "average" more frequently compared to the former group. This expectancy effect shows that judges actively try to preserve their judgmental degrees of freedom.

Alternatively, Bruine de Bruin and Keren's (2003) idea of unidirectional comparison processes also predicts increased judgment extremity. When stimuli possess unique positive features, judgments should become more positive, and when stimuli possess unique negative features, they should become more negative. Consequently, the resulting judgments should be a function of the comparison processes between adjacent stimuli. However, we observed such comparison (i.e., contrast) effects only in the beginning of the judgment sequence. The lack of comparison effects at position 6 does not fit with comparison process explanations in general. In addition, these authors observed similar effects for step-by-step and end-of-sequence. In our paradigm, serial position effects were all but eliminated by end-of-sequence judgments. Finally, the comparison framework would not predict or explain the expectancy effect observed by Fasold and colleagues (2012). Thus, we believe that the present data is best explained by a calibration process, which necessitates judgmental freedom preservation, resulting in extremity avoidance.

Threshold decisions

One might argue that the observed changes in grades are not particularly important. For poor performances, across Experiments 1–3, grades worsened only slightly from position 1 to position 5 ($M = -0.15$), although there is a substantial improvement for good performances ($M = 0.55$). However, even the small changes for poor exams might be of great importance. In many situations, categorical evaluative ratings involve threshold decisions, for example, to weed out the worst candidates, to select the best research proposals, or to accept or reject manuscripts submitted to scientific journals.

Calibration effects should be most pronounced for such threshold decisions. In the present case, we have one clear threshold and that is the grade 4.3, which means failing the exam. Symmetrically, we can define the grade 1.0 as a threshold for rewarding a distinguished performance. Considering both thresholds, we find clear calibration effects: Across Experiments 1–3, participants graded 17% of the observed performance as "fail" at position 1, while at position 5 they graded 31% as "fail" (a significant difference; Fisher's exact test $p = .013$). Conversely, participants used the "1.0" grade for 8% of the observed performances at position 5, but no performances at all (0%) received the best grade at position 1 (Fisher's exact test $p < .001$).

Thus, even if one considers the absolute changes on the grading scale as not important, it is difficult to argue with the increased chances to pass a threshold in later decisions, for the better or the worse (e.g., being accepted or being rejected).

Relation to social comparison models

Social comparison models might provide alternative explanations for these serial position effects (e.g., Mussweiler, 2003; Wänke et al., 2001). Participants might have become more extreme in their later judgments because they were able to compare the performance to contrasting exemplars. Across experiments, we used the average exams at positions 2 and 6 to test this possibility. Yet, we only observed contrast effects for the average exams in the beginning (position 2), and not in the end (position 6). This decreased influence of the preceding exams fits well with the idea of calibration, but speaks against comparison effects as an explanation for the more extreme judgments at position 5. Even more, as poor and good exam performances at position 5 were rated more extremely, one would expect stronger contrast effects for exam evaluations at position 6. However, across experiments, the effect of the preceding exams becomes negligible and insignificant later in the judgment series.

So how does the calibration idea relate to social comparison models in general? Let us illustrate the social comparison – calibration relation with a classic social comparison example about river lengths and whether a river is short, average, or long. For example, the Rhine (1233 km) is probably judged a large river in comparison to the Neckar (367 km), but in comparison to the Nile (about 6.600 km), it appears rather average. As stated above, we believe that such comparison processes are the means by which psychological ranges are established; however, as we have predicted and found here, contrast effects should diminish with the length of the judgment series. Thus, on a local level, contrast and exemplar comparison determine the outcome of categorical assessment; for example, given three categories of river sizes, "short – average – long", and the first river being the Nile and the second being the Rhine, the Rhine should be judged as short or average. If the first river is the Neckar and the second is the Rhine, the Rhine should be judged either as average or long, because the comparison outcome with the previous river determines this judgment, although the target is identical.

On a global level, calibration should determine the judgment. Having experienced the full range of possible river sizes in a series, the Rhine should be judged according to the mapping of the available categories to the observed range. Thus, the Rhine should be judged small when the series includes the Amazon (6448 km), the Mississippi (3778 km), and the Nile, even when the immediate river before judging the Rhine is the Hudson River (493 km). Conversely, if the series includes the Neckar, the Hudson, and the Thames (346 km), the Rhine should be judged large even if the preceding river is the Nile. With the length of the series, the local comparison influence should become weaker. This prediction coincides with the predictions from range-frequency theory (Parducci, 1965). And the fact that we observed no local contrast effects in Experiment 4 when judgments are made after observing the full sequences corroborates this interpretation.

Calibration as a process

The term calibration is often used in psychology to denominate the correspondence between stated confidence and rates of occurrence: "A judge is said to be calibrated if his or her probability judgments match the corresponding relative frequency of occurrence." (Liberman & Tversky, 1993, p. 162). Probably the most famous instance of calibration is the relation of confidence in judgments and judgment accuracy (Juslin, Winman, & Olsson, 2003; Lichtenstein, Fischhoff, & Phillips, 1982). Usually, experiments show that judges are badly calibrated, as their assessments concerning self-performance, eyewitness accuracy, or general

knowledge questions are far off from the real situation (Moore & Healy, 2008). We use the term with a different connotation – we refer to the process of mapping a perceived psychological range to an available category system (e.g., grades from *A* to *F*). Thus, rather than the outcome, we focus on the process and its implications.

This view is compatible with the existing use in the literature, but goes one step further and predicts *when* judges are well or poorly calibrated. For example, a common finding is that people have too much confidence in the accuracy of their answers to general knowledge questions (Fischhoff, Slovic, & Lichtenstein, 1977). Gigerenzer, Hoffrage, and Kleinbölting (1991) and in particular Juslin (1994) showed that overconfidence largely vanishes when the sample of question is drawn representatively from the universe of possible general knowledge questions. This translates directly to what we described as the psychological range on the relevant dimensions. If people have the chance to experience all levels of difficulty in trivia questions, their judgments will be better calibrated compared to when their experience is truncated; in our paradigm, when all students show very similar performances (e.g., all very poor, all very good, or all average).

Preventing serial position effects

The present results indicate a serious fairness problem for performances rated at early position within an evaluative judgment series. Especially good performances suffer if they appear first in a sequence. Consequently, an important question is how to prevent these serial position effects. In Experiment 4, we found end-of-sequence judgments to be successful; however, as discussed above, end-of-sequence judgments suffer from other problems, such as primacy and recency effects.

A solution could be a combination of both procedures: one might advise that judges observe the whole stimulus series and then judge each instance in a second observation round. While this procedure should prevent serial position effects theoretically, it is often not feasible practically. First, it is time-consuming, and second, it is impossible when ratings must be given immediately, for example in oral examinations or for performance evaluations in sport competitions.

Sport competitions directly prompt the reliance on expert judges as a solution – however, as shown in Experiment 3, experts are not immune to the proposed serial position effects. In addition, we found substantial calibration effects for expert referees in soccer (Memmert et al., 2008; Unkelbach & Memmert, 2008). And as stated in the introduction, experts might be particularly careful with extreme categories, especially for threshold decisions or when the judgment context is new, thereby showing the strongest calibration effects.

Another possibility is to provide judges with the chance to adjust initial ratings when sequences are over, which is clearly possible with written materials, as in Experiment 2. While this is an appealing idea, the large literature on anchoring and adjustment effects shows that once initial judgments have been made, adjustments are insufficient most of the time (e.g., Chapman & Johnson, 1999; Northcraft & Neale, 1987). At present, the most practical venue we are investigating is the *detachment* of a performance from the series, as this eliminates the need to preserve judgmental degrees of freedom. Consequently, if our theoretical reasoning is correct, this should also eliminate the observed calibration effects, especially for expert judges; they might be able to use an overarching psychological range for a given category system, without caring for the peculiarities of a specific judgment situation.

Conclusion

We showed that participants in the role of examiners evaluate good exams not as positive in the beginning of a series compared to the same exams later in the series. Conversely, poor exams were not judged as negative in the beginning as compared to the end. We predicted and explained this effect with the concept of calibration, combining the ideas of Parducci's (1965) range-frequency theory and Haubensak's (1992) consistency model: For series of categorical judgments, people must map the available category system to the psychological range of the stimulus series. To preserve judgmental degrees of freedom and thereby minimizing the possibility of significant consistency violations, they avoid extreme categories in the beginning. We believe that this calibration idea describes and explains many real-world phenomena and offers a theoretical starting point on how to make evaluations in serial judgments fairer and more accurate.

Acknowledgments

The present research was supported by a Frontier Grant from the University of Heidelberg to Christian Unkelbach and Daniel Memmert.

References

Bruine de Bruin, W. (2005). Save the last dance for me. Unwanted serial position effects in jury evaluations. *Acta Psychologica, 118*, 245–260.

Bruine de Bruin, W., & Keren, G. (2003). Order effects on judgments in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes, 92*, 91–101.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 79*, 1–39.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117–140.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 552–564.

Fasold, F., Memmert, D., & Unkelbach, C. (2012). Extreme judgments depend on the expectation of following judgments: A calibration analysis. *Psychology of Sport and Exercise, 13*, 197–200.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

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

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology, 19*, 323–340.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226–246.

Juslin, P., Winman, A., & Olsson, H. (2003). Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes, 92*(1–2), 34–51.

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.

Lewin, K. (1931). The conflict between Aristotelian and Galileian modes of thought in contemporary psychology. *Journal of General Psychology, 5*, 141–177.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin, 114*, 162–173.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York, NY: Cambridge University Press.

Martijn, C., Spears, R., Van der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology, 22*, 453–463.

Memmert, D., Unkelbach, C., Rechner, M., & Ertmer, J. (2008). Gelb oder kein Gelb? Persönliche Verwarnungen im Fußball als Kalibrierungsproblem [Yellow card or no yellow card? Soccer cautioning as a calibration problem]. *Zeitschrift für Sportpsychologie, 15*, 1–11.

Moore, D., & Healy, P. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502–517.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110*, 472–489.

Niedrich, R., Weathers, D., Hill, R., & Bell, D. (2009). Specifying price judgments with range–frequency theory in models of brand choice. *Journal of Marketing Research, 46*, 693–702.
Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*, 84–97.
Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407–418.
Parducci, A. (1968). The relativism of absolute judgments. *Scientific American, 219*, 84–89.
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.
Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105*, 143–146.
Scheer, J. K., & Ansorge, C. J. (1975). Effects of naturally induced judges' expectations on the ratings of physical performances. *Research Quarterly, 46*, 463–470.
Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*, 689–699.
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105.
Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*, 136–142.
Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect. *Current Directions in Psychological Science, 11*, 159–163.
Thompson, B. (2006). Research synthesis: Effect sizes. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583–603). Mahwah, NJ, US: Lawrence Erlbaum.
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.
Wänke, M., Bless, H., & Igou, E. (2001). Next to a star: Paling, shining, or both? Turning interexemplar contrast into interexemplar assimilation. *Personality and Social Psychology Bulletin, 27*, 14–29.
Wedell, D., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*, 341–356.
Wedell, D., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology, 102*, 233–248.