Psychology
2014. Vol.5, No.2, 91-98
Published Online February 2014 in SciRes (http://www.scirp.org/journal/psych)
http://dx.doi.org/10.4236/psych.2014.52014

Influences on the Marking of Examinations

Christina Bermeitinger[1], Benjamin Unger[2]
[1]Institute for Psychology, University of Hildesheim, Hildesheim, Germany
[2]Law firm Benjamin Unger, Hildesheim, Germany
Email: bermeitinger@uni-hildesheim.de

Received December 6th, 2013; revised January 5th, 2014; accepted February 3rd, 2014

In the present work, we examined a phenomenon highly relevant in the educational field for assessing or judging performance, that is, the question how the second examiner's marking is influenced by the evaluation of the first examiner. This phenomenon is known as anchoring in cognitive psychology. In general, in anchoring effects numeric information (i.e., the anchor) pulls estimations or judgments towards the anchor. One domain which is highly important in real life has been investigated only occasionally, that is, the marking of examinations. In three experiments, participants were asked to evaluate a written assignment. The mark (either good or bad) of a ficticious first examiner was used as the anchor. We found clear anchoring effects that were unaffected by feedback in a preceding task (positive, neutral, negative) or the expert status of the presumed first examiner. We discussed the problems related to this effect.

Keywords: Anchoring Effect; Marking; Feedback; Mood; Written Tests; Performance Judgments

Introduction

Our decisions and evaluations are influenced by many social and cognitive things and even numbers. Often, we are not aware of these influences, or we think we can shield our evaluations from such influences. However, there are a lot of psychological phenomena that demonstrate inadvertent influences on our cognition, for example, priming, framing, cueing, subliminal persuasion or advertising, and also anchoring.

Anchoring

The influence of numerical information on judgments is most often studied with the anchoring paradigm. Anchoring refers to the phenomenon that previously presented numerical information (i.e., the anchor) biases numerical judgments or estimates towards the anchor. Anchoring effects have been shown in a variety of different tasks and domains (for reviews see Chapman & Johnson, 2002; Epley, 2004; Furnham & Boo, 2011; Kudryavtsev & Cohen, 2010; Mussweiler, Englich, & Strack, 2004; Mussweiler & Strack, 1999), for example in probability estimation (Tversky & Kahneman, 1974), legal judgments (e.g., Englich & Mussweiler, 2001; Englich & Soder, 2009), price estimation (e.g., Englich, 2008), or general knowledge (e.g., Epley & Gilovich, 2001). Anchoring effects are present in a broad range of conditions—they result from implausible as well as plausible anchors (e.g., Mussweiler & Strack, 1999), from subliminal presentations of anchors (e.g., Mussweiler & Englich, 2005), when participants are forewarned or especially motivated not to be biased (e.g., Wilson, Houston, Etling, & Brekke, 1996), in experts and novices (e.g., Englich & Mussweiler, 2001), and in the laboratory as well as in real-world settings (e.g., Northcraft & Neale, 1987). Relevant but also irrelevant information that is clearly uninformative for the required judgment can serve as an anchor. For example, Northcraft and Neale (1987) tested anchoring effects in real estate agents. The authors provided a 10-page packet of information, which included a large amount of information regarding a piece of property currently for sale. The listing price for the property was varied as a relevant anchor which had an influence on the price the real estate agents had to state for the corresponding property. In contrast, Englich (2008) used an irrelevant anchor. She asked her participants to write down several numbers starting either with 10,150 or 29,150 before judging the price of a car. Such an irrelevant anchor had influences on the judgment as well. In the classic study of Tversky and Kahneman (1974), participants were required to estimate the percentage of African countries in the United Nations after spinning a wheel of fortune which determined the irrelevant anchor nevertheless had also influences on the estimation. Overall, anchoring effects are one of the most robust cognitive effects and, more specifically, cognitive heuristics.

Standard Anchoring vs. Basic Anchoring

Anchoring effects can be found with different approaches. The approaches most often applied are standard anchoring and basic anchoring. In standard anchoring (e.g., Tversky & Kahneman, 1974), participants are required to first make a comparative judgment (e.g., is the percentage of African countries in the United Nations higher or lower than 65%) and then an absolute judgment (e.g., an estimation of the percentage of African

countries in the United Nations). In this approach, participants must have a conscious representation of the numerical information given by the anchor (otherwise they could not answer the comparison task). In the basic anchoring approach, however, this is not necessary; no direct comparison is required. Participants are confronted with numerical information (either relevant or irrelevant, see above) and then they are asked to judge or estimate the object of interest. Both approaches lead to anchoring effects, but possibly due to different underlying processes (Englich, 2008; Mussweiler, 2002). In conclusion, the anchoring effect represents a robust phenomenon present in a broad range of diverse tasks and conditions.

Theories of Anchoring

There are different assumptions regarding the question why anchoring effects occur (e.g., Chapman & Johnson, 1999; Englich, 2008; Mussweiler, 2002; Mussweiler & Strack, 1999; Mussweiler et al., 2004). Originally, anchoring effects were explained in terms of insufficient adjustment from a given starting point (Tversky & Kahneman, 1974). In more recent views, anchoring effects, and specifically standard anchoring effects, are explained as the result of an active and elaborate hypothesis-testing process. It is assumed that persons hypothesize that the to-be-estimated "target" value is close to the anchor, and selectively activate information confirming this hypothesis. Thus, judges search for information consistent with their hypothesis via semantic associations between the target and the anchor. Such knowledge is activated, which implies a close relationship of target and anchor. Therefore, anchor-consistent knowledge is easier accessible than knowledge not consistent with the anchor. Furnham and Boo (2010) proposed that confirmatory search and selective accessibility are the main mechanisms contributing to the anchoring effect.

In contrast, basic anchoring effects are also explained by numeric priming—the anchor number is used as a reference point that is highly accessible simply due to the fact that it is one of the last numbers the person had in mind (for further perspectives on anchoring effects see e.g., Wegener, Petty, Detweiler-Bedell, & Jarvis, 2001; for review see e.g., Furnham & Boo, 2011).

Performance Judgments

Given the broad range of domains and conditions in which anchoring effects are present and investigated, it seems noteworthy that one domain has been subject to this research only occasionally despite its real-world relevance for people's careers—that is, performance judgment and specifically the marking of exams and assignments in school or university. A study by Dünnebier, Gräsel, and Krolak-Schwerdt (2009) is one rare exception. The authors investigated anchoring effects in teachers' assessment of written assignments depending on their expertise and the actual goal of assessment. The goal could be either to build a first impression or to give an educational recommendation. Results revealed an anchoring effect overall, but they were somewhat inconsistent with regard to the other factors. For example, only for a math test but not a German test did experts show a substantial anchoring effect with the processing goal of giving an educational recommendation.

Interestingly, in the domain of achievement judgments, there is often a consensus (especially in non-psychological areas) that examiners can make unbiased judgments. For example, in German jurisdiction, courts feel certain that an examiner usually is self-contained, self-dependent, free, unprejudiced, and completely unbiased by comments and marks of previous examiners (e.g., BVerwG, 2002). However, this ideal examiner is most likely far from reality (see also Brehm, 2003). Given this view of the idealized examiner, the lack of conclusive evidence available (cf. Dünnebier et al., 2009) is reason for concern. Thus, it seems desirable to have more conclusive results regarding biases and anchor effects in the specific context of marking.

The Present Experiments

In three experiments, we investigated anchoring effects on the marking of a written assignment. The written assignment comprised a question from the domain of motivational psychology. Our participants were undergraduate students from the University of Hildesheim, most of whom were enrolled in an introductory psychology unit on motivation and emotion. The written assignment was related to issues covered in the introductory unit. Thus, participants were more or less familiar with the topic of the written assignment. They were complete novices in the marking of examinations. However, these aspects also apply to several situations in real life—many examiners are rather inexperienced with examinations and sometimes even student assistants are instructed to pre-assess or mark written assignments. Additionally, in many cases—for example for German legal state examinations, for some school-leaving examinations in Germany, or in the example above in which student assistants mark the exams—examiners are confronted with assignments they did not build themselves. Further, they do not always teach the topics addressed in the written assignments. In summary, we tested the influence of anchors on the marking of examinations from student participants. Participants were non-experts, at least in the marking of examinations.

There is some evidence that non-experts show larger anchoring effects than experts. Chapman and Johnson (1994), for instance, found smaller anchoring effects for those participants who showed high certainty about their judgment. However, there are several studies that have demonstrated that evaluations by experts (e.g., experienced legal professionals, car experts, estate agents) are influenced by anchors as well (e.g., Englich & Soder, 2009; Mussweiler, Strack, & Pfeiffer, 2000; Northcraft & Neale, 1987). For example, Englich, Mussweiler, and Strack (2006) tested legal judges who were experts with extensive experience in the particular domain of law they were asked to judge during the study. These experts were influenced by randomly determined and irrelevant anchors to the same extent as judges who were experts in other domains (i.e., non-experts). In conclusion, in most cases expertise does not reduce the influence of an anchor. Some studies even found that only experts were influenced by an anchor (Englich & Soder, 2009). Thus, it seems adequate to test non-experts.

Our experiments are designed to closely resemble real situations in which written assignments have to be marked. We used the basic anchoring approach in which no comparison with the anchor is required. For a lot of written assignments (at least at German universities and for state examinations), it is typical that the second examiner knows the marking and evaluation of

the first examiner[1]. Thus, our participants also saw this information before individually marking the written assignments (for a similar procedure, see, e.g., Northcraft & Neale, 1987, who also provided a large amount of relevant information in addition to the anchor). Further, we used marks as the anchor, which clearly represents relevant information. Overall, it seems rather likely that we would find anchoring effects (that is, in general, we expected anchoring effects). However, it is important to actually show that performance judgments are influenced by the judgments of others, in particular given the myth of the ideal and unswayable examiner. To broaden our focus and to enlarge the innovative points of our study on anchoring effects in the context of performance judgments, we tested some further influences on this basic anchoring effect; first, whether a qualitative difference (i.e., the first examiner marking the assignment as "fail" vs. "pass") between the high and low anchor has an influence (Exp. 2 & 3); second, whether anchoring effects change when the first examiner is introduced as an expert vs. non-expert (Exp. 2); and third, whether positive, negative, or neutral feedback regarding participants' own performance in a preceding test affects the basic anchoring effect (Exp. 3).

Experiment 1

Method

Participants. The sample consisted of 49 students (45 female, 4 male) who were recruited from the introductory psychology unit on motivation and emotion at the University of Hildesheim. The median age was 21 years (ranging from 18 to 33 years). Subjects participated in several unrelated studies for course credit. They were randomly assigned to the conditions; each participant only marked one assignment.

Design. Experiment 1 was based on a one-factorial design. The factor anchor (high [3,0] vs. low [2,0]) was varied between participants.

Material. Essentially, the material consisted of the student's task (i.e. the exam question), the student's response, the report of the first examiner and his/her marking. The report included positive as well as negative aspects of the student's performance. The mark of the first examiner was either 2.0 or 3.0, which were both possible marks for the given performance.

Procedure. Participants were tested in groups of up to four persons, but participants worked individually in sound-attenuated chambers. All instructions were given on sheets of paper. First, participants were informed that we wanted to test different methods for evaluating student assignments and to find the best and fairest way for marking assignments. These points were emphasized to ensure that participants were motivated to participate and that they took the task seriously. Then, they were asked to work through the experimental materials in the given order and to read instructions carefully. Participants were informed that they had to judge the performance of a student in a written assignment, specifically regarding a question on the psychology of motivation. They were informed that they were the second examiner, that is, that a first examiner had already judged the performance and that they would see this judgment. Then, participants read the alleged student's task (i.e. the exam question) and they received some additional information, for example, they were given the marking scale (the standard marking scale at German universities: 1.0 "very good" – 1.3 – 1.7 – 2.0 "good" – 2.3 – 2.7 – 3.0 "satisfying" – 3.3 – 3.7 – 4.0 "sufficient" – ≥4.3 "failed") and some information on the points they should attend during their evaluation. Then, participants read the alleged student's response, which was presented on two pages. On the next page, participants saw the report of the alleged first examiner and his/her marking. Subsequently, participants were asked: "How do you as a second examiner mark the written assignment?" Participants were additionally asked to write down the main arguments why they gave this mark for the given performance.

Results and Discussion

As Figure 1 reveals, participants' judgments were influenced by the given anchor. Participants who were confronted with the higher anchor (i.e., 3.0) gave higher marks than those who were confronted with the lower anchor (i.e., 2.0) producing a significant main effect of anchor in a one-way ANOVA, $F(1, 47) = 7.36$, $p = .009$, $\eta_p^2 = .14$.

The result showed that the general task, including materials, procedure, and participants, produced clear anchoring effects. The results also showed that overall, participants tended to mark the written assignment worse than intended: With the lower anchor of 2.0, participants marked the assignment with $M = 2.91$ which is close to 3.0 (i.e., the higher anchor which led to a mark of $M = 3.48$).

Experiment 2

Participants in Experiment 1 tended to evaluate the given assignment as worse than "good" (i.e. >2.0). Thus, in Experiment 2, we used higher anchors (i.e. 3.7 vs. 4.3). The higher anchor (4.3) indicates a mark that is associated with a "fail". Note that in Experiment 1 only 3 out of 25 participants of the

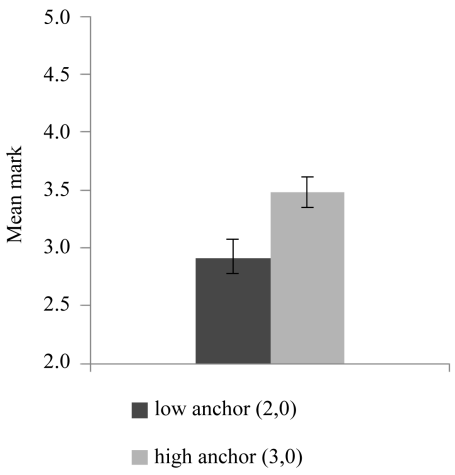

Figure 1.
Mean mark by anchor (high [3,0] vs. low [2,0]). Error bars represent the standard error of the mean.

[1]One anonymous reviewer during the reviewing processes of this paper wondered about the relevance of our study: "In actual scoring no one in their right mind would allow the score from a different scorer to be shown. Therefore, this paper has created a completely artificial setting that bears no relationship to applied reality." Actually and unfortunately, that is not the case. We created the situation for our participants as closely as possibly to the applied reality as we found it in various exam situations, at least in Germany.

low anchor condition, and only 2 out of 24 participants of the high anchor condition, evaluated the student's assignment as "fail". This implies that a higher anchor is not associated with higher "fail" rates per se. Thus, in Experiment 2 we used 3.7 as the lower anchor and 4.3 as the higher anchor—there was hence not only a quantitative difference between both anchors but also a qualitative difference. We expected that participants who were confronted with the higher anchor more often marked the assignment as "fail" than participants who were confronted with the lower anchor. Additionally, we varied whether the first examiner was introduced as an expert with many years of experience in psychology and in evaluating psychological exams or as a non-expert (i.e., a student of informatics without any experience in psychology). The question was whether participants were more influenced by the anchor from the first examiner who was introduced as an expert. Although previous research (e.g., Englich & Mussweiler, 2001; Englich et al., 2006) could not find differences in the anchoring effect dependent on anchor relevance, it might be that in the present case the informational value of the anchor for the given task (which should be higher if the first examiner is introduced as an expert rather than a non-expert) does play a role.

Method

Participants. The sample consisted of 76 undergraduate students (63 female, 13 male), again recruited from the introductory psychology unit on motivation and emotion at the University of Hildesheim, none of whom had participated in Experiment 1. The median age was 21 years (ranging from 18 to 36 years). Subjects participated in several unrelated studies for course credit. They were randomly assigned to the conditions.

Design. Experiment 2 was based on a 2 (anchor: high [4,3] vs. low [3,7]) × 2 (first examiner: expert vs. non-expert) design. Both factors were varied between participants.

Material and Procedure. Materials and the procedure were the same as in Experiment 1 with the following exceptions. First, participants were additionally informed that the first examiner was either an expert with many years of experience in psychology and in evaluating psychological exams or that the first examiner was another student (i.e., a first-year informatics student) with no experience in psychology or evaluating exams. Second, while the report of the first examiner who was introduced as an expert was the same as that used in Experiment 1, with very few minor changes, the report of the first examiner who was introduced as a non-expert was shorter and more colloquial. The report of the non-expert included positive as well as negative aspects of the student's performance, too. Third, the mark of the first examiner was either 3.7 or 4.3 (i.e., a "fail").

Results and Discussion

Mean marks (see **Figure 2**) were subjected to a 2 (anchor: high vs. low) × 2 (first examiner: expert vs. non-expert) analysis of variance (ANOVA). There was a significant main effect of anchor, $F(1, 72) = 8.43$, $p < .01$, $\eta_p^2 = .11$. On average, participants who were confronted with the higher anchor (i.e., 4.3) gave higher marks than those who were confronted with the lower anchor (i.e., 3.7). In contrast, the main effect of first examiner and the interaction of anchor and first examiner were not significant, both Fs < 1, ps > .85, indicating that it did not matter whether the first examiner was introduced as an expert or non-expert (see also **Figure 2**).

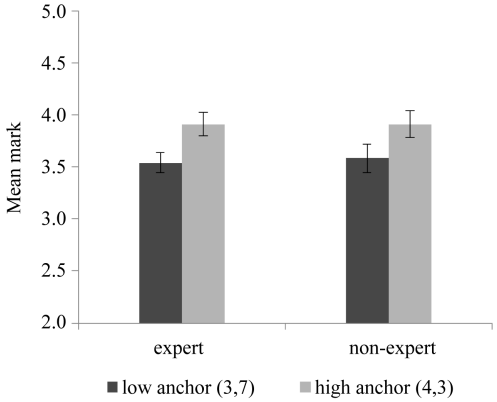

Figure 2.
Mean mark by anchor (high [4,3] vs. low [3,7]) and first examiner (expert vs. non-expert). Error bars represent the standard error of the mean.

Overall, participants of the high anchor condition (4.3, i.e., "fail") marked the assignment with $M = 3.91$ (i.e., on average, the student would have passed). In this condition, there was no significant difference ($p = .14$) between the number of participants who evaluated the assignment as failed ($n = 14$) and the number of participants who evaluated the assignment as passed ($n = 24$). That is, a substantial number of participants evaluated the assignemt as "fail". In contrast, participants of the low anchor condition (3.7) marked the assignment on average with $M = 3.56$. In this condition, only 5 participants marked the assignment as "fail"; there were significantly more participants who marked the assignment as "pass" ($n = 33$, $p < .001$).

Again, we found clear anchoring effects. Additionally, we found evidence that participants who were confronted with the "fail" anchor more often marked the assignment as "fail" than participants who were confronted with the lower anchor, as there was no difference between "fail" and "pass" marks in the high anchor condition but there were significantly more "pass" marks in the low anchor condition. However, the "fail" anchor did not lead to more "fail" than "pass" marks but only increased the proportion of "fail" marks.

Additionally, we varied whether the first examiner was introduced as a non-expert or as an expert with many years of experience in psychology and in evaluating psychological exams. For this factor, we found no effect—it played no role whether the first examiner was an expert or not. This finding matches with previous research, in which even irrelevant information completely uninformative for the given task (e.g., Critcher & Gilovich, 2008; Englich, 2008; Englich et al., 2006; Tversky & Kahneman, 1974) caused robust anchoring effects of the same magnitude as anchoring effects from relevant anchors (e.g., Englich & Mussweiler, 2001; Englich et al., 2006). That is, in the domain of marking of examinations—highly relevant for the individual's career—robust anchoring effects were found which were independent of whether the anchor represented highly valuable (i.e., the anchor was given by an expert) or less valuable (i.e., the anchor was given by a non-expert) information. Typically, such findings are explained by assuming that the accessibility of anchor-consistent information biases judgments or estimations independent of the informational relevance of the anchor (e.g., Furnham & Boo, 2011).

Experiment 3

Experiment 1 and 2 revealed clear anchor effects on the marking of a written assignment. In Experiment 3, we mainly tested the influence of positive, negative, or neutral (fictitious) feedback regarding participants' own performance in a preceding task on these anchoring effects. Such feedback could affect the processing of following information by mechanisms also operational in affective or semantic priming (e.g., Clore, Wyer, Dienes, Gasper, Grohm, & Isbell, 2001; Klauer & Musch, 2003; Neely, 1991). In the given situation, feedback may be able to influence one's mood. Such feedback is a very interesting factor in the context of marking, first, because some guides recommend examiners not to evaluate assignments or exams when they are in a (too) bad or (too) good mood, and second, mood can be influenced in everyday life by a lot of things that are outside one's control, for example by feedback in social or achievement situations. Thus, in Experiment 3 we manipulated whether participants got positive, negative, or neutral feedback regarding their own performance in a computer test. Typically, the own affective state is used as source of information (cf., "mood as information" hypothesis, for review see e.g., Schwarz & Clore, 2003) when people had to evaluate something. Schwarz and Clore assumed (and showed) that own states are misread as a response to the object people had to judge. As a result, they evaluate things more favorable under positive rather than negative states (and vice versa). Thus, affective information can affect decision making indirectly by influencing how we process information (e.g., Clore et al., 2001; Englich & Soder, 2009; Schwarz, 2001). While in a happy mood, a more holistic and heuristic processing style is typical, whereas a more elaborate and critical information processing style is often associated with negative/sad mood (for an overview see e.g., Huntsinger, Clore, & Bar-Anan, 2010).

However, results from anchoring studies seem to show a different pattern regarding their dependence on mood. Englich and Soder (2009) tested the influence of mood on judgments (Study 1: in a legal shoplifting case; Study 2: in ordinary estimates). For non-experts, the authors only found anchoring effects when participants were in a sad mood but not when participants were in a happy mood. In contrast, for experts, Englich and Soder either found no anchoring effects at all (Study 2) or anchoring effects occurred no matter what mood participants were in. Based on these results, we expected anchoring effects at least after negative feedback (see also Bodenhausen, Gabriel, & Lineberger, 2000). Additionally, we added a condition in which participants were first examiners, that is, without any anchor from the mark of another examiner.

Method

Participants. The sample consisted of 79 undergraduate students (68 female, 11 male), again recruited from the introductory psychology unit on motivation and emotion at the University of Hildesheim or on campus. None of these had participated in one of the other experiments. The median age was 22 years (ranging from 18 to 42 years). Subjects participated in several unrelated studies either for course credit or remuneration. They were randomly assigned to the conditions.

Design. Experiment 3 was based on a 3 (anchor: high [4,3] vs. low [2,7] vs. no anchor) × 3 (feedback: positive vs. negative vs. neutral) design. Both factors were varied between participants.

Material and Procedure. Materials and the procedure were identical to Experiment 1 with the following exceptions. First, we introduced a control condition in which no anchor and no report of another examiner was shown. In this condition, participants were informed that they were the first examiner and that the assignment would be subsequently judged by a second examiner who would see their report and their mark. Second, in the conditions in which participants were second examiners, the mark (i.e., the anchor) of the first examiner was either 2.7 or 4.3 (i.e., "fail"), that is, there was again a qualitative difference between both anchors (as in Experiment 2). Third, we manipulated which (fake) feedback on their own performance in a preceding computer task (in which we recorded reaction times and error rates) was given to the participants.

This computer task was introduced as a reliable measure of general mental efficiency (including intelligence, processing speed, and so on). The computer task was run using E-Prime software (version 1.3) with standard PCs and 17-in. CRT monitors. Instructions were given on screen. One to three stimuli occurred simultaneously at nine possible locations of an invisible 3 × 3 grid (with grid points located at 75%, 50%, and 25% positions of the vertical and horizontal full screen span). The stimuli were either squares or dots and either of blue, yellow, green, or red color. The background color was white. The computer task comprised 100 trials, each lasting 800 ms. However, single stimuli could be presented for one, two, three, or four sequential trials. Thus, the duration of one stimulus could be 800, 1600, 2400, or 3200 ms. The participants' task was very simple: They were instructed to press the space key as fast as possible whenever a yellow square or a blue dot appeared anywhere on the screen, which was the case in 22 out of the 100 trials. However, it was rather difficult to supervise the whole field and stimuli appeared rather fast. Thus, it was not possible to get a good appraisal of one's own performance. The whole computer task took approximately 2 minutes. Most importantly, at the end, participants were informed that they had achieved a mean reaction time for their correct responses of 547 ms. Participants of the positive feedback condition were additionally informed that they had achieved a very good result, which only 10% of comparable participants had also managed to achieve. Participants of the negative feedback condition were informed that they had achieved a below-average result, and that more than 70% of comparable participants had achieved a better result. Participants of the neutral feedback condition received no feedback[2]. Thereafter, participants proceeded to the marking task, and were fully debriefed at the end of the experiment.

[2]In a pre-test with 46 student participants, we tested whether the chosen feedback was adequate to induce different mood states. In the pre-test, participants worked through the same reaction time task as used for the main experiment and got the same fictitious feedback (negative, neutral, or positive) regarding their performance. Thereafter, they worked through the Mehrdimensionaler Befindlichkeitsfragebogen (multidimensional mental state questionnaire, Steyer, Schwenkmezger, Notz, & Eid, 1997). Therein, mood was measured with eight adjectives (after recoding: the higher the score, the better the mood). Results showed that participants' mood was indeed influenced by the given feedback, $F(1, 43) = 4.66$, $p = .02$. Participants who got positive feedback had significantly higher mood scores than participants who got neutral ($t(28) = 2.83$, $p = .01$) or negative ($t(28) = 2.64$, $p = .01$) feedback. (There were no significant differences between the mood scores of participants who got negative vs. neutral feedback, $t < 1$, $p > .77$.). That is, the task was indeed adequate to induce differences in mood, at least between positive and negative/neutral feedback.

Results and Discussion

Mean marks (see **Figure 3**) were subjected to a 3 (anchor: high vs. low vs. no) × 3 (feedback: positive vs. negative vs. neutral) analysis of variance (ANOVA). There was a significant main effect of anchor, $F(2, 70) = 16.32$, $p < .001$, $\eta_p^2 = .32$. In contrast, the main effect of feedback and the interaction of anchor and feedback were not significant, both Fs ≤ 1, ps > .36, indicating the feedback participants got had no effect (see also **Figure 3**).

On average, participants who were confronted with the higher anchor (i.e., 4.3) gave higher marks than those who were confronted with the lower anchor (i.e., 2.7), $t(50, 42.94) = 5.85$, $p < .001$ (t-test for unequal variances), $M_{\text{high anchor}} = 4.00$, $SD = 0.46$, $M_{\text{low anchor}} = 3.08$, $SD = 0.65$. Additionally, participants who were confronted with the higher anchor (i.e., 4.3) gave higher marks than those who were confronted with no anchor, $t(52, 44.23) = 4.60$, $p < .001$ (t-test for unequal variances), $M_{\text{no anchor}} = 3.24$, $SD = 0.73$. In contrast, participants who were confronted with the lower anchor did not differ significantly from participants who were confronted with no anchor, $t(50) = 4.30$, $p = .43$.

As before, in the high anchor condition (4.3, i.e., "fail") there was no significant difference ($p = .70$) between the number of participants who evaluated the assignment as failed ($n = 12$) and the number of participants who evaluated the assignment as passed ($n = 15$). That is, a substantial number of participants evaluated the assignment as "fail". In contrast, only 2 participants of the low anchor condition (i.e., 2.7) and only 5 participants of the no anchor condition marked the assignment as "fail"; there were significantly more participants who marked the assignment as "pass" ($n = 23$, $p < .001$ and $n = 23$, $p = .001$, for the low anchor condition and no anchor condition, respectively).

Again, we were able to show an anchoring effect and we could replicate the results from Experiment 2 regarding the pattern of "fail" vs. "pass" marks depending on the presence of a "fail" or "pass" anchor. Further, we found a mean mark of 3.24 for participants who were not confronted with any anchor. Participants of the no anchor condition did not differ significantly from participants of the low anchor condition. Combined with results from Experiment 1 (especially the condition in which we used an anchor of 2.0), we found evidence that the higher anchor in the given task was able to bias evaluation towards a higher (i.e., worse) mark. However, the lower anchor did not bias the marking significantly towards a better mark (i.e., in Exp. 1 an anchor of 2.0 led to a mean mark of 2.9; in Exp. 3 an anchor of 2.7 lead to a mean mark of 3.08—both results are close to the no anchor condition with a mean mark of 3.24; please note that the distance between the mean unbiased mark and the lowest anchor in Exp. 1 was 1.24 and the distance between the mean unbiased mark and the highest anchor was 1.06; however, there was no large bias towards the low anchor). This result seems highly important for the given context as it might be especially problematic that a second examiner is influenced by the bad mark of a first examiner, which might lead to disproportionately high numbers of evaluations in which the given mark would be worse than the appropriate mark.

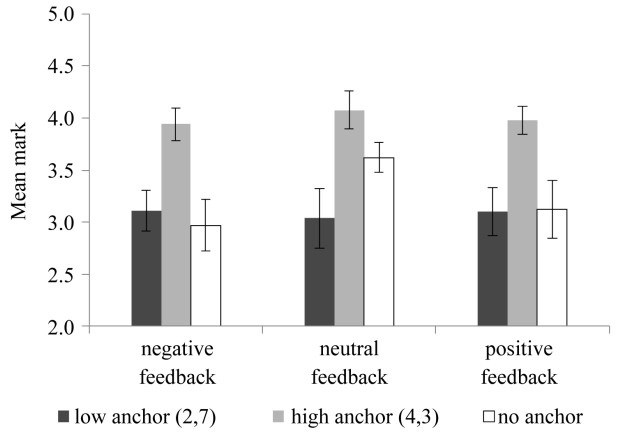

Figure 3.
Mean mark by anchor (high [4,3] vs. low [2,7] vs. no) and feedback (positive vs. negative vs. neutral). Error bars represent the standard error of the mean.

In contrast, we found no influence of the (fictitious) feedback in the preceding computer task, which was either negative, positive, or neutral. Here, we can only speculate why we did not find an effect of this manipulation. First, the manipulation may have not been adequate to affect participants differently in the main experiment. Either there was no influence or at least no differential influence, that is, it might be that the computer test per se was, for example, frustrating (comparable to the sad mood condition with non-experts of Englich & Soder, 2009). However, if we assume that mood or emotion, and hence also the effects of our computer-task and feedback, was used as source of information (Schwarz & Clore, 2003), we would expect worse marking in general—compared to our previous experiments. This was not the case. We found comparable results, especially for anchors that were identical across experiments (i.e., 4.3 in Experiment 2 and Experiment 3). Second, perhaps the manipulation was initially successful, however, it may have lasted too short a time to influence the marking or the bias at the end of the assignment. Third, there may actually be no influence of feedback (or own affective state) on the marking of examinations—at least in the tested situation. Fourth, the manipulation may have indeed been successful, however, as in Englich and Soder (2009, Study 1) for experts, anchoring effects occurred no matter which feedback participants got or in which mood they were. It has to be noted that we did not check whether participants' mood was influenced in the main experiment.

General Discussion

In three experiments, we found conclusive evidence for anchoring effects in the marking of examinations. However, we found no evidence for the impact of feedback on a preceding task, or whether the anchor was introduced as an expert evaluation or an evaluation of a novice. There are several implications which we will discuss in the following.

As could be shown in Experiment 3 in comparison with Experiment 1, a bad mark in particular was able to influence participants' marking (however, note that in Experiment 3 there was also a qualitative difference between the anchors). This seems to be unrelated to the specific number of the anchor: in the study of Dünnebier et al. (2009), it was also the bad mark that influenced participants' judgments more (at least for the German assignment) although "bad" in Dünnebier et al. was associated with a *lower* anchor (they used a 0 - 15 grade scale with 15 as the best grade). Participants of the no anchor condi-

tion did not differ significantly from participants of the low anchor condition. That is, an anchor representing a bad mark influences examiners more towards a bad mark. In contrast, an anchor representing a good mark does not influence examiners to the same extent towards a good mark. In real marking situations this seems rather alarming. If a first examiner gives an assignment a bad mark, it seems rather hard for second examiners to evaluate the assignment in an unbiased fashion. Especially in these cases, examinees may be evaluated more harshly than their actual performance would warrant. Thus, it is more likely overall that assignments marked by two examiners are evaluated too harshly rather than too leniently.

Given the apparent robustness of this effect, we need to consider different procedures for marking examinations. Critically, even strategies or manipulations aiming to reduce anchoring effects have typically failed to eliminate the effect entirely (e.g., Mussweiler et al., 2000). Mussweiler et al. (2000, Study 1) asked car experts to estimate the price of a car. The experts were given an anchor but then also some anchor-inconsistent information ("[Someone] mentioned yesterday that he thought this value is too high/low"). They were then confronted with the question "What would you say argues against this price?" By doing so, Mussweiler et al. were able to reduce the anchoring effect significantly compared to a condition in which no anchor-inconsistent information was provided. However, there was still a trend of an anchoring effect.

In the same vein, Mussweiler (2002) found larger anchoring effects with a similarity focus compared to a difference focus. In this study, participants were first required to list either as many similarities between two visual scenes as they could find or as many differences as they could find. This task was introduced as being completely independent of the following (anchoring) task on general knowledge. To transfer such effects to the marking of examinations, perhaps it would be helpful if second examiners should have the specific focus of finding arguments against the marking of the first examiner. In contrast, in the current practice, second examiners are virtually rewarded when they agree with the marking of the first examiner—they often have to justify a deviating evaluation but they can easily write "I fully agree with the first examiner" (which, of course, is much less effort) when they award the same mark.

In the context of achievement judgment by teachers, Dünnebier et al. (2009) showed that there was no significant anchoring effect for experts with the processing goal of giving an educational recommendation (at least in their German assignment). For judgments of examinations at university or single assignments in state examinations, it has to be assumed that examiners do not necessarily have that goal, for example, because the single assignment represents just one amongst others. Additionally, examiners cannot know how important the single assignment will be for the future of the examinee. Thus, it seems unlikely that they evaluate assignments with the specific goal of giving an educational recommendation—or in a broader sense with the goal to provide their most exact judgment on the given assignment. Perhaps, it might be helpful to emphasize that the single assignment may be highly relevant for a student's career, and an adequate assessment will provide highly valuable information (and potentially serve as a recommendation) for potential employers.

Additionally, it might be discussed whether examinations should be evaluated independently by two examiners without knowledge of each other's marking. Such a procedure would of course require safeguards such that it is not undermined by verbal agreements between examiners. Anchoring effects still occur when persons are trained or even informed about the influence of anchors. Thus, pure training of reviewers seems insufficient to reduce anchoring effects. However, it might be interesting to see results from future research directly concerned with this question in the context of marking. Furthermore, the use of automated scoring might be a method which could circumvent the problems of biased marking (for an overview see Shermis & Burstein, 2003). However, potential advantages and disadvantages (e.g., greater time requirements in the case of independent reviewing or the lack of human interaction in the case of automated scoring) have to be balanced.

Last but not least, we want to point out some limitations of our experiments which may stimulate some future studies. First of all, we tested student participants. Anyway, further research could investigate effects in expert examiners. Second, one might ask whether other mood induction techniques might be better suited to influence anchoring effects (or whether more participants are needed to find differences). Additionally, it could be helpful to test participants' mood also before, during, and after the main task (i.e., judging). Third, we only tested relevant anchors. It might be interesting whether anchors completely unrelated to the judgment task (e.g., as given in the classic study by Tversky & Kahneman, 1974) also influence the judgment of examinations. Additionally and generally, it would be interesting to relate the topic of anchoring effects and findings of rater biases. Fourth, we used a rather unstandardized task (which is used in a lot of exams, of course). However, it might be interesting to test also more standardized tasks and answers—are judgments of more standardized tasks influenced comparably by the mark of a first examiner? Fifth, which influence has the relationship between first and second examiner? Sixth, our participants had to judge only one answer. It would be very interesting to have a situation in which one participant has to judge a lot of answers which might create also reference points across different answers and allow comparative judgments. Eighth, it would be interesting to investigate the influence of rewards, for example for rapid judgments. In this context, the report of the second examiner is most often rather short and in applied reality it is very short in cases in which the second examiner gives the same mark as the first examiner. Generally, in reality, fast judgments are rewarded.

Conclusion

In conclusion, we have shown that anchoring effects are also found in the domain of achievement judgments. We tested students who were rather unfamiliar with marking situations, but this situation is not uncommon at universities, where complete novices are often given the task of evaluation of assignments. From literature, we pointed out several ways to potentially reduce such anchoring effects—for example, a difference focus, blind marking, emphasizing the importance of correct evaluations, or automated scoring. It remains to be seen whether such conditions could be implemented and whether they are instrumental for fairer and more objective evaluations of students' achievements. In sum, we cannot ignore the discrepancy between our results, showing clear anchoring effects, and the ideal of the unprejudiced examiner, which is actually far from reality.

Acknowledgements

We thank Nicolas Salzer, Luise Maier, Laura Flatau, David Eckert, Lena Zepter, and Elke Förster-Fröhlich for their help in data collection. We thank Ullrich Ecker for improving the readability of this article.

REFERENCES

Blankenship, K. L., Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *Journal of Experimental Social Psychology, 44*, 1465-1476. http://dx.doi.org/10.1016/j.jesp.2008.07.005

Bodenhausen, G. V., Gabriel, S., & Lineberger, M. (2000). Sadness and susceptibility to judgmental bias: The case of anchoring. *Psychological Science, 11*, 320-323. http://dx.doi.org/10.1111/1467-9280.00263

Brehm, R. (2003). The human is unique also as examiner. *Neue Juristische Wochenschrift, 56*, 2808-2810.

BVerwG [Federal Administrative court of Germany] (2003). Urteil vom 10.10.2002-6 C 7/02. *Neue Juristische Wochenschrift, 56*, 1063-1064.

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making, 7*, 223-242. http://dx.doi.org/10.1002/bdm.3960070402

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 19*, 115-153. http://dx.doi.org/10.1006/obhd.1999.2841

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120-138). New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511808098.008

Clore, G. L., Wyer, R. S., Dienes, B., Gasper, K., Gohm, C., & Isbell, L. (2001). Affective feelings as feedback: Some cognitive consequences. In L. L. Martin, & G. L. Clore (Eds.), *Theories of mood and cognition: A user's guidebook*. Mahwa, New Jersey: Lawrence Erlbaum.

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making, 21*, 241-251. http://dx.doi.org/10.1002/bdm.586

Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Biases in teachers' assessments of student performance: An experimental study of anchoring effects. *Zeitschrift für Pädagogische Psychologie, 23*, 187-195. http://dx.doi.org/10.1024/1010-0652.23.34.187

Englich, B. (2008). When knowledge matters: Differential effects of available knowledge in standard and basic anchoring tasks. *European Journal of Social Psychology, 38*, 896-904. http://dx.doi.org/10.1002/ejsp.479

Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the court-room. *Journal of Applied Social Psychology, 31*, 1535-1551. http://dx.doi.org/10.1111/j.1559-1816.2001.tb02687.x

Englich, B., & Soder, K. (2009). Moody experts: How mood and expertise influence judgmental anchoring. *Judgment and Decision Making, 4*, 41-50.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin, 32*, 188-200. http://dx.doi.org/10.1177/0146167205282152

Epley, N. (2004). A tale of tuned decks? Anchoring as accessibility and anchoring as adjustment. In D. J. Koehler, & N. Harvey (Eds.), *The Blackwell handbook of judgment and decision making* (pp. 240-256). Oxford: Blackwell Publishers. http://dx.doi.org/10.1002/9780470752937.ch12

Epley, N., & Gilovich, T., (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*, 391-396. http://dx.doi.org/10.1111/1467-9280.00372

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*, 35-42. http://dx.doi.org/10.1016/j.socec.2010.10.008

Huntsinger, J. R., Clore, G. L., & Bar-Anan, Y. (2010). Mood and global-local focus: Priming a local focus reverses the link between mood and global-local processing. *Emotion, 20*, 722-726. http://dx.doi.org/10.1037/a0019356

Klauer, K. C. & Musch, J. (2003). Affective priming: Findings and theories. In J. Musch, & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion*. Mahwah, NJ: Lawrence Erlbaum.

Kudryavtsev, A., & Cohen, G. (2010). Illusion of relevance: Anchoring in economic and financial knowledge. *International Journal of Economic Research, 1*, 86-101.

Mussweiler, T. (2002). The malleability of anchoring effects. *Experimental Psychology, 49*, 67-72. http://dx.doi.org/10.1027//1618-3169.49.1.67

Mussweiler, T., & Englich, B. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes, 98*, 133-143. http://dx.doi.org/10.1016/j.obhdp.2004.12.002

Mussweiler, T., & Strack, F. (1999). Comparing is believing: A selective accessibility model of judgmental anchoring. *European Review of Social Psychology, 10*, 135-167. http://dx.doi.org/10.1080/14792779943000044

Mussweiler, T., Englich, B., & Strack, F. (2004). Anchoring effect. In R. Pohl (Ed.), *Cognitive illusions: A handbook of fallacies and biases in thinking, judgement, and memory* (pp. 183-200). London, UK: Psychology Press.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*, 1142-1150. http://dx.doi.org/10.1177/01461672002611010

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Erlbaum.

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. http://dx.doi.org/10.1016/0749-5978(87)90046-X

Schwarz, N. (2001). Feelings as information: Implications for affective influences on information processing. In L. L. Martin, & G. L. Clore (Eds.), *Theories of mood and cognition: A user's guidebook* (pp. 159-176). Mahwa, New Jersey: Lawrence Erlbaum.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychology Inquiry, 14*, 296-303.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah: Lawrence Erlbaum.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *The multidimensional mental state questionnaire: Manual*. Göttingen: Hogrefe.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1130. http://dx.doi.org/10.1126/science.185.4157.1124

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010). Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Journal of Consumer Psychology, 20*, 5-16. http://dx.doi.org/10.1016/j.jcps.2009.12.003

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology, 37*, 62-69. http://dx.doi.org/10.1006/jesp.2000.1431

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125*, 387-402. http://dx.doi.org/10.1037/0096-3445.125.4.387