View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by AIS Electronic Library (AISeL)

Can Experience be Trusted? Investigating the Effect of Experience on Decision Biases in Crowdworking Platforms

Thomas Görzen
Paderborn University, Germany
Thomas.Goerzen@upb.de

Abstract

Companies increasingly involve the crowd for collective decision making and, to aggregate the decisions, they commonly average the scores. By ignoring crowdworkers' different levels of experience and decision biases, this method may not favor the best outcome. Alternatively, decisions can be weighted in favor of the more experienced judges in the crowd. However, previous research is inconclusive as to whether more experienced individuals are any better at avoiding decision biases. To answer this question, we conduct online crowd-based experiments with a range of treatments, comparing the anchoring effect of individuals with different levels of experience. Results indicate that not only does greater experience not protect crowdworkers from the anchoring effect but it increases their confidence in their decision, compared to less experienced individuals, even if they are wrong. Our findings provide valuable insights for both researchers and practitioners interested in improving the effectiveness of crowdworking decision-making.

1. Introduction [1]

Companies using crowdworking for idea generation often face the challenge of having to screen hundreds or thousands ideas submitted by the crowd. How do they select the more valuable ones out of a vast number of ideas? For example, when the car manufacturer Fiat turned to the crowd to suggest ideas for its new Fiat 500 model, the call generated over 170,000 design ideas and more than 20,000 comments on specific aspects, such as the shape of the exhaust or of the chrome bumpers [1]. Since most companies do not have the resources to evaluate such a large number of ideas, or only with disproportionate effort, the crowd is increasingly used not only to generate new ideas but also to evaluate them, thus doubling up the challenge of how to tackle the number of ideas generated and evaluated. Nevertheless the method of using a large number of judgements - the so-called wisdom of the crowd – continues to offer key advantages by i) maximizing the amount and the diversity of information by drawing on a large number of people from a wide range of information backgrounds and, ii) reducing the potential impact of 'outliers' - extreme decisions based on unreliable or simply inaccurate information sources. For a crowd to be wise, however, it has to meet the following conditions: i) it must be diverse, ii) decentralized, and iii) independent in its judgement [3]. The downside of the wisdom of the crowd, as the literature also suggests, is that individuals might be biased in their decision-making. Based on insights from cognitive psychology, biased decision making, often simply labeled as "decision biases", can be described as "[…] *human behavior which goes beyond the rationality assumptions of neoclassical and new institutional economic theory*" [4]. Examples of decision biases include individuals failing to adequately judge probabilities, making wrong predictions or being too extreme in their judgement [5].

Another prominent bias that might occur in the context of idea evaluation by an online-crowd is the well-established anchoring effect [6]. This effect describes a disproportionate influence that an initially presented value has on a decision maker [6]. The reasons for the occurrence of this bias in this context are as follows: First, online platforms are characterized by increasing information richness and often provide information such as ratings given by other workers ([7], [8]), which could act as an anchor [9]. Second, especially in organic crowdworking campaigns where the judgement process is structured into two or more phases, the risk of the anchoring effect occurring is quite high when the information which potentially acts as anchor is revealed to workers in subsequent stages [10]. Lastly, the anchoring effect occurs for both types of tasks, objective (e.g. estimating the height of the Eiffel

[1] Note that [2] draws on the same research environment and preliminary results of an earlier version have been presented as a poster at the Multikonferenz Wirtschaftsinformatik (MKWI) 2018.

URI: https://hdl.handle.net/10125/59876
ISBN: 978-0-9981331-2-6
(CC BY-NC-ND 4.0)

tower) and subjective estimations (evaluating the creativity of an idea) [11], which covers the majority of typical task designs on crowdworking platforms.

Regardless of this potential risk of biased decision making, currently the favorite method for using the wisdom of the crowd approach is to simply average the judgements of all individual decisions in the crowd [12], even though this can lead to suboptimal outcomes neglecting, as it does, external information such as experience. As an alternative to simply averaging individual judgements, researchers (e.g. [13]) have proposed weighted models that favor more experienced judges in the crowd. The assumption behind this approach is that more experienced judges are less likely to be affected by the anchoring effect or, if they are affected, that their estimation will still be more valid than that of a less experienced person. However, previous studies on decision making in the offline context show contrasting results. While some studies indicate that knowledgeable people are less influenced by biases (e.g. [14]), others show that even knowledgeable people with experience in a given context are significantly biased (e.g. [15]). With respect to these conflicting results, we want to shed light on this question and aim to answer the following research question:

Are more experienced people in the crowd less prone to the anchoring effect in their decision-making?

To answer this question, we conducted experiments on a commercial crowdworking platform, with the crowd being given 80 different business model ideas to evaluate. We implement an anchor, i.e. information about the previous rating results given by others, using it as a treatment, and analyze its effect on individual raters' decisions. After idea evaluation, we asked each participant about her background experience, using several dimensions of experience relevant to the task, such as business models, product category, their experience in retail and how long they have been a member of the platform. In addition, after have completed their evaluation, raters were asked how confident they felt about their ratings. Results indicate that experience in different dimensions does not decrease the probability to follow an anchor and therefore does not protect against biased decision making. Moreover, experienced judges are more confident about their decision, even if they are wrong – in our case, deviant from experts' ratings, which we used as a benchmark.

Our study makes contributions to extant research in several ways: (1) in contrast to most previous studies, our research on the anchoring effect is conducted in the context of the large and still growing field of crowdworking, following the proposed call for additional research on cognitive biases in the IS context [8]; (2) we extend prior research on the anchoring effect by taking into account not only one but several dimensions of experience (product, context, etc.); (3) we use a subjective evaluation task in contrast to objective tasks such as general knowledge questions; (4) we investigate the robustness of weighted models to aggregate multiple opinions that favor more experienced judges in the crowd. This allows us to investigate whether these models could be used to provide valid results, even if potentially biased decisions might occur within the crowd. Lastly, our analysis offers valuable insights for practitioners, by helping task designers in crowdworking to avoid potential pitfalls.

2. Related Literature and Hypotheses

2.1. Anchoring Effect in Decision Making

Because of humans' bounded cognitive resources [16], if humans would strictly adhere to the laws of logic and probability, even a single complex everyday situation would require more computations than can be performed in a human lifetime [17]. Therefore, people tend to apply fast but fallible heuristics in their daily life to reduce the complexity of their decision making. The downside of this, however, are cognitive biases like the anchoring effect. The anchoring effect describes the disproportionate influence of an initially presented value on decision makers [6]. This cognitive bias is subconscious and involves noticing an initial value or starting point (the anchor), which influences one's decision (subconsciously) by adjusting one's response in a direction that seems appropriate.

The main underlying mechanisms of anchoring are selective accessibility [18] and confirmatory search [19]. Selective accessibility in this case means that as long as an initially presented anchor lies within the boundaries of a known category (for example: the height of the Eiffel tower is about 300 meters and not 3,000), participants construct a mental model that selectively increases the accessibility of anchor-consistent information [18]. In line with that, confirmatory search means that when a plausible anchor is presented, people focus on activating information that is consistent with this value and neglect information that deviates from the anchor [19]. The occurrence of the anchoring effect has been shown in a variety of different domains such as general knowledge [20] or probability estimation [19]. Previous studies show that both uninformative anchors and anchors with informational relevance to the task itself are able to influence a decision. In their classic study, [6] randomly generated anchor values were obtained by spinning a wheel of

fortune between 0 and 100 and afterwards asking the participants to estimate the percentage of African countries in the United Nations. The given anchor values had a strong influence on their estimation since higher anchors significantly increased the estimations compared to lower ones [6]. Another study found that the estimation of an athlete's performance could be anchored by the number on his jersey [21]. Anchors with relevance to the task can also lead to the anchoring effect: in an example from the legal domain, higher damage awards were obtained when higher compensations were requested in court [22]. Following previous results, we assume that the anchoring effect will generally affect a crowd's decision making when an anchor is displayed. More specifically, when comparing the distribution of ratings with and without displaying an anchor, we expect both results of the evaluation to differ. Hence, we derive the following hypothesis:

Hypothesis 1: The distribution of ratings will differ depending on whether an anchor is displayed.

2.2. Influence of Experience on Anchoring Effect

Experience in the context of judges who evaluate ideas with respect to creativity is defined as "[…] *the acquisition and cumulative knowledge of reality, mechanisms, rules, and procedures related to a specific domain*" [23]. Previous literature on the influence of experience on the anchoring effect shows contradictory results. On the one hand, it suggests that experienced people utilize information in their decision making differently to those less experienced than them ([24], [25]) since they:

- process information more quickly due to practice and skill
- process information in their domain of expertise more efficiently
- know more than others and can access that knowledge better
- are less likely to be influenced by factors that could influence information processing and capacity allocation.

Hence, people with higher expertise should arguably be less influenced by anchors. The assumption that greater expertise and experience protects against the anchoring effect to a certain extent is supported by previous studies which assert that both people with high certainty about their answer [19] and those with greater relevant knowledge are less influenced by anchors [14].
Table 1 presents several studies of the influence of experience on the occurrence of the anchoring effect.

Table 1. Studies on the influence of experience on the anchoring effect

Article	Context	Experienced less influenced?
[14]	General knowledge questions	✓
[19]	Lotteries	(✓)
[26]	Estimating the value of houses	-
[27]	Judgements of event probabilities	-
[28]	Fraud estimation and critical event prediction	-
[18]	Evaluating the value of a cars	-
[15]	Hypothetical crime case	-

On the other hand, results from other studies indicate that an anchor does not only affect inexperienced decision makers but also experts. For example, car experts (dealers and mechanics) with all the necessary information available were influenced by anchors when evaluating the value of a car [18], and legal professionals by information irrelevant to the case [15]. Regarding the latter, results of an experiment with a hypothetical crime case show that judges' sentencing decisions were strongly influenced by the prosecutor's demand. The sentences given for the same hypothetical crime case were significantly higher for participants who were presented with a high sentencing demand than for those with a low demand [15].

Based on these results it could be argued that the underlying mechanisms for the anchoring effect are so engrained in fundamental cognitive processes that people regardless of their level of experience and knowledge could fall victim to this effect. This potentially invalidates the notion that anchors should only affect inexperienced decision makers rather than expert participants. Based on this argument and the majority of results in the field that demonstrate that anchoring has a robust effect on human decision making regardless of the experience of decision makers, we derive our second hypothesis:

Hypothesis 2: Higher experience does not decrease the probability to follow a displayed anchor.

3. Research Methodology

To test our hypotheses, we design an experiment which enables us to (1) analyze the occurrence of the anchoring effect in idea evaluation on a crowdworking platform and (2) investigate whether experience might protect people from being influenced by the anchoring effect. In the context of crowdworking, idea evaluation tasks for new products, services or business models represent a typical task [9]. The ideas in our experiment were taken from previous research [29], in which students generated business model ideas for perfume in a classroom experiment. After having been given basic knowledge about business models and the Business Model Canvas [30], participants generated business model ideas for perfume. Only the best ideas, self-selected by each participant, were collected. All ideas are presented in the same way, consisting of the nine elements of the Business Model Canvas [30]. We then designed an evaluation task on Crowdflower, a commercial crowdworking platform which draws on potential contributors distributed worldwide. Because it would be unreasonable to ask each participant to evaluate all 80 business model ideas, we divided the ideas into eight blocks of ten ideas each and randomly assigned each business model idea to one block. Each participant had to rate the displayed ideas on a seven-point scale in terms of creativity, novelty and usefulness [31]. This part of our experiment represents the control condition in which each individual had to evaluate the ideas without encountering an anchor. We further designed two additional experiments with two different treatments, enabling us to investigate the anchoring effect. First, we designed an experiment (Crowd Anchor) where the information about the previous evaluation resulting from the control condition of each idea was displayed above the rating scale. Second, we designed an additional experiment (Random Anchor) where the only difference was that for the displayed rating, each idea was assigned a randomly generated rating between 1.0 and 7.0. The general task design of the control condition was retained for both additional experiments. Figure 1 shows a screenshot of a rating scale in one of the treatment conditions with the anchor displayed in the left-hand corner above the rating scale.

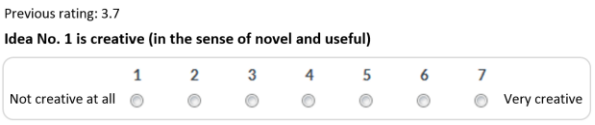

Figure 1. Treatment with displayed anchor

To investigate whether we find support for our first hypothesis, i.e. the general occurrence of the anchoring effect, we compare the average ratings of ideas for each experimental condition. Further, we also analyze the rating distributions of the different experimental conditions to see whether the display of an anchor influences the overall rating distribution as well as the potential deviation of ratings from experts' evaluation. Finally, we consider the influence of experience on anchoring. The experimental design is illustrated in Figure 2.

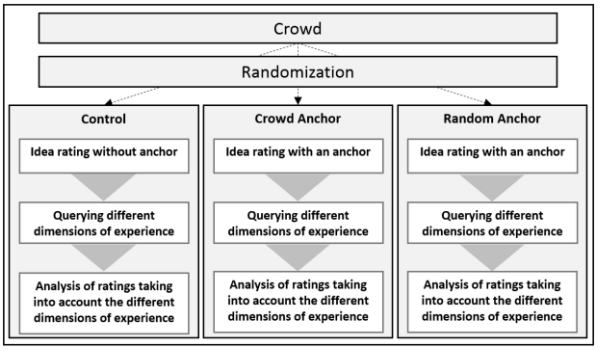

Figure 2. Experimental design

To avoid the occurrence of a learning effect and potential rating bias through users who assign themselves to several tasks in succession, we allowed each participant to evaluate only one block. Each block of ideas was evaluated by twenty different contributors, each earning 0.50$ for the evaluation of one block (= ten ideas). After idea evaluation was completed, we pre-screened the rating of each idea and excluded all those with a standard deviation below 0.5, assuming that below this threshold, participants wanted to earn money as easily as possible and merely clicked through the task quickly. In sum we analyzed 4,560 individual ratings, 1,490 for the control condition and 3,070 for both treatment conditions.

To analyze the influence of experience, all participants had to complete a short survey in addition to the evaluation task. The aim of the survey was to collect a range of data to assess participants' experience or prior knowledge across the following dimensions: Context/market mechanism, platform experience, product knowledge, and business models. More specifically, we asked participants about the length of their membership on the platform in number of months (platform experience), their experience in retail in number of months (context/market mechanism), their knowledge about perfume (four items, product knowledge), and their experience of business models (two items). After the idea evaluation was completed, we further asked each subject to rate their confidence in evaluating the business model ideas presented to them. All scales ranged from 1 = "Not at all" to 7 = "Absolutely agree". Additionally, to find out whether the randomly assigned rating of an idea strongly

deviates from a more professional evaluation of idea quality, we recruited two experts who would serve as a benchmark [32]. One came from our university's entrepreneurship center and the other works as a senior in-house consultant in a large corporation.

4. Results

4.1. Analyzing the General Occurrence of an Anchoring Effect

We first calculated the average rating of each idea in each experimental condition. In general, the descriptive statistics (Table 2) of the different conditions do not differ much from each other.

Table 2. Descriptive statistics

Condition	Obs.	Mean	Std. Dev.	Min	Max
No Anchor	80	4.58	.412	2.60	5.87
Crowd Anchor	80	4.59	.479	3.35	5.50
Random Anchor	80	4.39	.745	2.60	5.87

To further investigate the different ratings of each idea in the different conditions, we compare their average ratings, using parametric tests. First, we check whether rating distributions show any significant deviation from a normal distribution by using the Kolmogorov-Smirnov test nor the Shapiro-Wilk test. However, results of both tests show no significant deviation from a normal distribution (lowest p-value: .128). Next, we test for homogeneity of variance for the average rating of ideas for each experimental condition by conducting a Levene's test based on the mean as well as the median. Both results based on the mean $F(2, 237) = 19.27$, $p < .001$ and median $F(2, 237) = 19.04$, $p < .001$ show significant results, implying that the homogeneity of variances has been violated. Therefore, we conducted t-tests specified for data with unequal variances. Following [33], we additionally report the effect size for each experimental condition. Table 3 presents results of pairwise comparisons of the average ratings for each experimental condition. First, comparing average ratings for the condition where no anchor was displayed with the condition where the average rating resulting from the control condition of each idea was displayed above the rating scale, shows no significant difference in ratings. However, this result is not surprising since we did not expect the display of the average rating from the control condition to significantly change the rating for the first experimental condition (Crowd Anchor).

Table 3. Comparison of average ratings

	Mean	Std. Dev.	Diff-erence	Effect Size (r)
No Anchor	4.584	.412	-.006 (.462)	r = .007
Crowd Anchor	4.591	.479		
No Anchor	4.584	.412	.187** (.025)	r = .154
Random Anchor	4.397	.745		
Crowd Anchor	4.591	.479	.193** (.026)	r = .153
Random Anchor	4.397	.745		

*Note: *p <.10; **p < .05; ***p< .01*

Moreover, we expected both ratings to lead to comparable results because the occurrence of an anchoring effect in this case would mean that participants incorporate the displayed anchor, based on the previous decisions from the control condition, into their decision. Comparing the random anchor with both the crowd anchor and the control condition shows statistically significant differences. Hence, we conclude the following: first, the comparison between the control condition and the situation where a randomly generated rating is displayed demonstrates that the occurrence of an anchor affects the subsequent rating. Second, the type of anchor displayed also plays a role since comparing both treatment conditions with different anchors leads to a statistically significant difference in rating results. To further investigate the general occurrence of the anchoring effect, we also analyzed the distribution of individual ratings from each experimental condition (Figure 3). In contrast to the analysis of the average rating above, we now compare the rating distribution of each individual participant in each condition, i.e. 1,490 for the control condition and 3,070 for both treatment conditions (out of a total of 4,560 individual ratings). As can be seen, the given distributions differ depending on whether and which anchor was shown in the condition. First, we can see that in our first treatment condition (Crowd Anchor) the range of ratings has been reduced, while in our second treatment condition (Random Anchor), the range of ratings has increased when a randomly generated anchor was shown.

This result is in line with the previous comparison of average ratings above, supporting our interpretation that the occurrence of an anchor effect in this case decreases the variance of ratings, leading to comparable evaluation results.

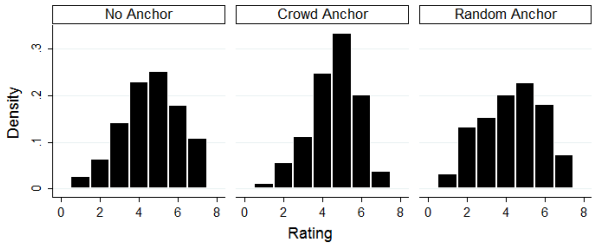

Figure 3. Distribution of ratings

This visual inspection supports our result that participants incorporate the anchor provided and moreover, that the specific value of the displayed anchor strongly influences subsequent ratings.

In addition to a visual inspection of these distributions, we compare the two rating distributions by using the Kolmogorov-Smirnov statistic. Our first hypothesis is supported if the distributions significantly differ from each other. Results are reported in Table 4.

Table 4. K-S statistics for comparison of distributions

Conditions	Difference (Combined K-S)
No Anchor - Crowd Anchor	.0716 (.001)***
No Anchor - Random Anchor	.0851 (< .001)***
Crowd Anchor - Random Anchor	.1378 (< .001)***

*Note: *p <.10; **p < .05; ***p< .01*

Results indicate that the distribution of ratings in all experimental conditions differ significantly from each other. In addition, we investigate the potential negative effect of anchoring, i.e. the tendency of decision makers to make decisions towards a presented value that deviates from the experts' assessment. To investigate this effect, we calculated the difference from each rating to the experts' rating and compared the differences of crowd ratings to experts' ratings for both experimental conditions.

Table 5. Comparison of differences between crowds' rating and experts' rating

	Mean	Std. Err.	Diff.
Crowd Anchor	1.329	.0255	-.2481*** (.0395)
Random Anchor	1.577	.0300	

*Note: *p <.10; **p < .05; ***p< .01*

Results (Table 5) show that the randomly generated anchor (Random Anchor) leads to a statistically significant ($p < .001$) higher distance to the experts' ratings than the displayed anchor from the control condition (Crowd Anchor). This result highlights the potential negative effect of anchoring, since initially displayed wrong values (anchors) can significantly increase the distance to the actual desired result.

Accordingly, we find support for hypothesis 1 since the presence of an anchor changes the distribution of ratings, implying the occurrence of an anchoring effect in crowd decisions. Moreover, the anchoring effect is robust for the anchors displayed in the different experimental conditions.

4.2. Analyzing the Anchoring Effect in Relation to Level of Experience

To find out whether more experienced people are less prone to be influenced by an anchor, we investigate whether participants followed it, additionally factoring in the influence of participants' experience. To answer our research question, we investigate the data from our two treatment conditions to see whether participants followed the displayed anchor. We used a logit regression with the dependent variable = 1 if the person's rating was consistent with the displayed anchor. Due to the fact that people could only rate in integers (for example: 4) while the anchor was provided as a float (for example: 4.2), the dummy variable is = 1 if the person's rating was closer to the provided anchor. If the anchor was 4.4, for example, and the person rated the idea with 4, the dummy variable would be 1. In contrast, if the provided anchor was 4.6, the dummy variable would be 1 if the person rated the idea with 5 since this value is closer to the provided anchor than 4. To avoid multicollinearity in our model, since some variables for experience consisted of several items, we conducted a principal component analysis [34] to summarize multi-item variables. Thus, we consider the following model [35]:

$$Y_j^* = \beta_0 + \beta_1\, Mem_j + \beta_2\, BMC\_Exp_j + \beta_3\, Perf\_Exp_j + \beta_4\, Retail\_Exp_j + \beta_5\, Conf\_Rating_j + \varepsilon_j,\ Y = 1[Y_j^* > 0].$$

Mem_j represents the reported length of membership on the platform of a person, BMC_Exp_j captures the multi-items for the reported experience in business models, while $Perf\_Exp_j$ captures the different items for the reported experience and knowledge of perfume. $Retail\_Exp_j$ represents the reported experience in months of a participant working in retail, while $Conf\_Rating_j$ covers the reported confidence of the participant when evaluating the business model ideas.

Further, we use clustered robust standard errors on the participant level to account for autocorrelation in the data since each participant rated ten different ideas [35]. Results (Table 6) indicate that experience in different dimensions, such as relevant to the key product (perfume), using and evaluating business models or having worked in retail, does not significantly decrease the probability of following the displayed anchor.

Table 6. Results of logit regression

Variable	Coef.	Std. Err.[2]	z	P > |z|
Mem_j	-.0004	.0031	-.013	.893
BMC_Exp_j	-.0670	.0530	-1.26	.206
$Perf\_Exp_j$	.0617	.0385	1.60	.109
$Retail\_Exp_j$	-.0007	.0008	-.089	.375
$Conf\_Rating_j$	.0576	.0524	1.10	.272
Constant	-1.212***	.2807	-4.32	.000
Controls[1]	✓			

*Notes: *p < .10; **p < .05; ***p < .01; [1]Gender and age; [2]clustered robust std. err. by participant*

However, one might argue that especially when analyzing the data from the condition where the average rating from the control condition was displayed (Crowd Anchor), it might be rational for participants to follow the displayed average rating from the control condition since their rating matches the average rating from the control condition. To address this point, we again used our logit model mentioned above and solely analyzed the dataset from our second treatment condition (Random Anchor). Since we displayed randomly generated ratings as anchors for each idea, we assume that the probability that these anchors match the quality of ideas or reflects the professionals' rating of the idea (= experts' rating) is quite low. Further, we only analyze ideas where the difference between the displayed anchor and the experts' rating was ≥ 2. In sum, we analyzed 740 observations. Results are illustrated in Table 7.

In line with previous results, the estimators indicate that experience does not significantly decrease the probability to follow the displayed anchor, which applies even when the displayed anchor strongly deviates from the quality of the ideas (difference ≥ 2 from experts' rating). These results are constant for different dimensions of experience. Thus, our results support our second hypothesis. In other words, a higher level of experience does not increase protection against making a biased decision by following a randomly generated anchor.

Table 7. Results of logit regression for random anchors with difference to experts' rating ≥ 2

Variable	Coef.	Std. Err.[2]	z	P > |z|
Mem_j	.0096	.0066	1.45	.147
BMC_Exp_j	-.0423	.1278	-.33	.740
$Perf\_Exp_j$	-.0029	.0880	-.03	.974
$Retail\_Exp_j$	.0001	.0018	.08	.937
$Conf\_Rating_j$	.2292*	.1380	1.66	.097
Constant	-2.859***	.7665	-3.73	.000
Controls[1]	✓			

*Notes: *p < .10; **p < .05; ***p < .01; [1]Gender and age; [2]clustered robust std. err. by participant*

Surprisingly, our results indicate that people who follow the displayed anchor in this situation feel more confident about their rating. However, we assume that higher confidence does not increase the probability to follow the anchor. Rather, we suspect a reverse effect. More specifically, we assume that participants who follow the anchor feel more confident about their rating. Hence, when participants in our experiment rated the idea as suggested by the displayed anchor, they felt more confident about their decision since it seemed to be in line with others.

In addition to investigating whether participants follow the anchor displayed by using a dummy variable, we further used a continuous variable to analyze the potential influence of different dimensions of experience on the occurrence of the anchoring effect. Here we calculated the difference between the rating submitted by each participant and the anchor displayed for all participants in our second treatment condition (Random Anchor). If the difference was negative (for example: 2.4 [anchor displayed] – 4.0 [participants' rating]) a positive value was calculated by multiplying the value by minus one to avoid that negative and positive values cancel each other out. We used these differences as our continuous, dependent, variable and the different dimensions of experience as independent variables, and conducted a linear regression. If coefficients of different dimensions for experience are statistically significant and positive it would suggest that experience does protect against the occurrence of the anchoring effect.

Results (Table 8) indicate, however, that experience does not significantly increase the distance between the displayed anchor and the submitted rating, suggesting that the more experienced still follow the anchor. In contrast, both membership (in months) on the platform and experience of using the business model canvas seems to decrease the distance between the displayed anchor and the submitted rating.

Table 8. Results of linear regression for random anchors

Variable	Coef.	Std. Err. [2]	t	P > |t|
Mem_j	-.0069**	.0027	-2.52	.013
BMC_Exp_j	-.0893*	.0480	-1.86	.065
$Perf\_Exp_j$	.0740	.0496	1.49	.137
$Retail\_Exp_j$	-.0003	.0005	-.65	.514
$Conf\_Rating_j$	.0518	.0636	.82	.416
Constant	1.345***	.3479	3.87	.000
Controls[1]	✓			

*Notes: *p < .10; **p < .05; ***p < .01; [1]Gender and age; [2]clustered robust std. err. by participant*

Further, we used the difference between the displayed anchor and the submitted rating as continuous variable and conducted a linear regression for the second treatment condition (Random Anchor) where the difference between the displayed anchor and the experts' rating was ≥ 2. Hence, we analyzed a situation in which we assume that the probability of the displayed anchors matching the quality of ideas or reflecting the professionals' rating of the idea (= experts' rating) is quite low.

Table 9. Results of linear regression for random anchors with difference to experts' rating ≥ 2

Variable	Coef.	Std. Err. [2]	t	P > |t|
Mem_j	-.0116***	.0040	-2.83	.005
BMC_Exp_j	-.0678	.0666	-1.02	.310
$Perf\_Exp_j$	.0841	.0733	1.15	.253
$Retail\_Exp_j$	-.0002	.0007	-.30	.762
$Conf\_Rating_j$	.0742	.0844	.88	.381
Constant	1.583***	.4900	3.23	.002
Controls[1]	✓			

*Notes: *p < .10; **p < .05; ***p < .01; [1]Gender and age; [2]clustered robust std. err. by participant*

In line with previous results, the coefficients indicate that experience does not significantly increase the distance between the displayed anchor and the submitted rating (Table 9). Again, the opposite is the case since membership (in months) on the platform seems to decrease the distance between the randomly generated anchor and the submitted rating of participants.

Next, we further investigate the influence of the length of platform membership on the confidence of ratings since previous results indicate that (i) participants who follow the displayed anchor feel more confident about their decision (Table 7), and (ii) that membership on the platform (Table 9) decreases the distance between a randomly generated anchor and participants' rating. Therefore, we want to answer the question of whether more experienced people feel more confident in the assessment, even if they are wrong.

Based on our previous results and in line with the literature, (e.g. [15]) which used experience working on a job as a proxy for experience, we use a participant's length of platform membership as a proxy of experience. We differentiate between two groups: experienced (membership >=12 months) and relatively inexperienced participants (membership <12 months). Further, we define a "wrong decision" as a submitted rating which deviates at least 2 points from the experts' evaluation. We used a two-sided t-test to compare the confidence in being wrong (as defined above) for both groups. In sum, we analyzed 936 observations from both experimental conditions. Results (Table 10) show that more experienced participants are statistically significantly (p = .002) more confident about their rating even if this rating is wrong in the sense that it strongly deviates from experts' rating.

Table 10. Participants' confidence in being wrong

Membership	Mean	Std.Err.	Diff.
<12 months	5.00	.0437	-.1881***
>=12 months	5.18	.0509	(.0671)

*Notes: *p < .10; **p < .05; ***p < .01; Non-parametric test leads to qualitatively comparable results.*

5. Conclusion

The currently most commonly used application to aggregate multiple judgements in crowdworking consists of simply averaging individual judgements, which entails several disadvantages and is susceptible to systematically ignores biased decisions. Hence, to overcome these problems researchers have proposed weighted models that favor more experienced judges in the crowd. However, this approach assumes that more experienced people are less inclined to making biased decisions.

Therefore, we investigate whether more experienced people in a crowd are less prone to decision biases. By conducting several experiments on a crowdworking platform, using the established anchoring effect as a treatment [6], we aim to shed light on this question. While previous literature on the influence of experience on the anchoring effect shows contradictory results, our results indicate that experience in different dimensions does not decrease the probability of following an anchor and therefore does not protect against biased decision-making. This result is consistent across different anchors. In addition, experienced members in the crowd

feel more confident about their decision than less experienced persons, even when they are wrong (i.e. deviate from the experts' evaluation of idea quality).

We contribute to the body of literature on mechanisms of aggregating multiple judgements in crowdworking. In contrast to the previous literature in this context which compared the absolute results of different aggregating mechanisms, we specifically investigate the potential of the occurrence of biased decisions for a weighted aggregation mechanism. We also contribute to the literature on the anchoring effect, specifically in respect of the influence of experience on the occurrence of this effect. Further, in contrast to previous research in the offline context which mainly focused on one dimension of experience (e.g. number of years working in a specific job), we take several dimensions of experience into account and investigate their respective influence on anchoring. Our results also carry managerial implications.

First, companies who want to use the crowd for idea evaluation should be aware that even experienced members of the crowd might be influenced by anchors. Hence, weighted mechanisms to aggregate multiple judgements should be chosen carefully since this mechanism can also lead to biased results. Second, from the perspective of crowdworking platform designers, especially where the judgement process is divided into two or more steps, our results highlight that displaying the information (e.g. judgements results) from previous process steps in the following steps should be considered with caution, if biases are to be avoided.

While our study provides important contributions, we also acknowledge certain limitations. First, self-reported experience might not be an objective measure since participants might over-or underestimate their own experience. However, we argue that self-reporting to collect information about participants' experience is a common practice in experimental research and also used in several other studies (e.g. [36]). Further, even if over-or underestimation in self-reported experience might have occurred, this should not lead to a systematic difference between the participants in the different experimental conditions. Second, the task assigned to the crowd (i.e. evaluation of business models for perfume) might be quite specific. However, since we conducted several experiments on a commercial crowdworking platform, our task design had to appear natural in this context since the crowd usually solves similar kinds of tasks. Finally, we did not measure experience with regards to having knowledge of biases, e.g. whether participants are aware of these and try to avoid them. However, we suggest that additional studies involving other ideas or tasks should be conducted to investigate whether the effect is constant for different tasks or types of ideas. Future work could seek to gain additional insight into the reasons for biased decision making. For example, using the "think aloud method" (e.g. [37]) by asking subjects in a laboratory setting to talk through their idea evaluation could help to understand the cognitive processes involved when solving the given tasks, enhancing our understanding of decision biases.

6. References

[1] Kleemann, F., Voß, G. G., and Rieder, K. "Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing." Science, Technology & Innovation Studies (4:1), 2008, pp. 5-26.

[2] Görzen, T. "Trust the Experienced? Investigating the Effect of Experience on Decision Making in the Crowd." Proceedings of the Multikonferenz Wirtschaftsinformatik, Research-in-Progress, 2018, pp. 989-995.

[3] Surowiecki, J. "The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations." Doubleday Books, 2005, New York.

[4] Carter, C. R., Kaufmann, L., and Michel, A. "Behavioral Supply Management: a Taxonomy of Judgment and Decision-Making Biases". International Journal of Physical Distribution & Logistics Management (37:8), 2007, pp.631-669.

[5] Gilovich, T., Griffin, D. and Kahneman, D. "Heuristics and Biases: The Psychology of Intuitive Judgment." 2002, Cambridge University Press, Cambridge, UK.

[6] Tversky, A. and Kahneman, D. "Judgment Under Uncertainty: Heuristics and Biases." Science 185, 1974, pp. 1124-1131.

[7] Duan, W., Gu, B., and Whinston, A. B. "Informational Cascades and Software Adoption on the Internet: An Empirical Investigation". MIS Quarterly (33:1), 2009, pp. 23-48.

[8] Fleischmann, M., Amirpur, M., Benlian, A,, and Hess, T. "Cognitive Biases in Information Systems Research: A Scientometric Analysis. In: Proceedings of the 22nd European Conference on Information Systems, 2014, pp. 1-21

[9] Chiu, C.-M., Liang, T.-P. and Turban, E. "What can Crowdsourcing do for Decision Support?" Decision Support Systems 65, 2014, pp. 40-49.

[10] Eickhoff, C. "Cognitive Biases in Crowdsourcing." Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM), 2018.

[11] Orr, D. and Guthrie, C. "Anchoring, Information, Expertise, and Negotiation: New Insights from Meta-Analysis". Ohio State Journal on Dispute Resolution (21:3), 2006, pp. 597–628.

[12] Larrick, R. P., Mannes A.E. and Soll, J. B. "The Social Psychology of the Wisdom of Crowds." Krueger, J.I., ed. Frontiers of Social Psychology: Social Judgment and Decision Making, 2011, pp. 227-242.

[13] Aspinall, W. "A Route to more Tractable Expert Advice." Nature (463:7279), 2013, pp. 294-295.

[14] Wilson, T. D., Houston, C. E., Etling, K. M., and Brekke, N. "A New Look at Anchoring Effects: Basic Anchoring and its Antecedents." Journal of Experimental Psychology: General 125, 1996, pp. 387-402.

[15] Englich, B. and Mussweiler, T. "Sentencing under Uncertainty: Anchoring Effects in the Courtroom." Journal of Applied Social Psychology 31, 2001, pp. 1535-1551.

[16] Griffiths, T. L., Lieder, F., and Goodman, N. D. "Rational use of Cognitive Resources: Levels of Analysis between the Computational and the Algorithmic." Topics in Cognitive Science (7:2), 2015, pp. 217–229.

[17] Lieder, F., Griffiths, T. L., Huys, Q. J. M., and Goodman, N. D. "The Anchoring Bias reflects rational use of Cognitive Resources. Psychonomic Bulletin & Review (25), 2018, pp. 322–349.

[18] Mussweiler, T., Strack, F., and Pfeiffer, T. "Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility." Personality and Social Psychology Bulletin 26, 2000, pp. 1142-1150.

[19] Chapman, G. B., and Johnson, E. J. "The Limits of Anchoring." Journal of Behavioral Decision Making 7, 1994, pp. 223-242.

[20] McElroy, T. and Dowd, K. "Susceptibility to Anchoring Effects: How Openness-to-Experience Influences Responses to Anchoring Cues." Judgment and Decision Making 2, 2007, pp. 48-53

[21] Critcher, C. R. and Gilovich, T. "Incidental Environmental Anchors." Journal of Behavioral Decision Making 21, 2008, pp. 241-251.

[22] Marti, M. W., and Wissler, R. L. "Be Careful What You Ask For: the Effects of Anchors on Personal Injury Damages Awards." Journal of Experimental Psychology: Applied 6, 2000, pp. 91-103.

[23] Galati, F. "Complexity of Judgement: What Makes Possible the Convergence of Expert and Nonexpert Ratings in Assessing Creativity." Creativity Research Journal (27:1), 2015, pp. 24-30.

[24] Garrett, S. K., Caldwell, B. S., Harris, B. S. and Gonzalez, M. C.. "Six Dimensions of Expertise: A more Comprehensive Definition of Cognitive Expertise for Team Coordination." Theoretical Issues in Ergonomics Science (10:2), 2009, pp. 93-105.

[25] Hoffman, R. R., Feltovich, P. J. and Ford, K. M. "A General Framework for Conceiving Expertise and Expert Systems in Context." In: P.J. Feltovich, K.M. Ford and R.R. Hoffman, eds. Expertise in context. Menlo Park, CA: AAAI Press, 1997, pp. 543-580.

[26] Northcraft, G. B., and Neale, M. A. "Experts, Amateurs, and Real Estate: An Anchoring and Adjustment Perspective on Property Pricing Decisions." Organizational Behavior and Human Decision Processes 39, 1987, pp. 84-97.

[27] Wright, W. F. and Anderson, U. "Effects of Situation Familiarity and Financial Incentives on Use of the Anchoring and Adjustment Heuristics for Probability Assessment." Organizational Behavior and Human Decision Processes 44, 1989, pp. 68-82.

[28] Joyce, E. J., and Biddle, G. C. "Anchoring and Adjustment in Probabilistic Inferences in Auditing." Journal of Accounting Research 19, 1981, pp. 120-145.

[29] John, T. and Kundisch, D. "Why Use the Canvas for Idea Generation? A Design Theory and First Evidence." Working Paper, 2015, Paderborn University.

[30] Osterwalder, A. and Pigneur, Y. "Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers." 2010, Hoboken, NJ: Wiley.

[31] Amabile, T. M. "Creativity in Context: Update to the Social Psychology of Creativity." 1996, Boulder, CO: Westview Press.

[32] Bayus, B. L. "Crowdsourcing New Product Ideas Over Time: An Analysis of the Dell IdeaStorm Community." Management Science (59:1), 2013, pp. 226-244.

[33] Lin, M., Lucas Jr., H.C., and Shmueli, G. "Research Commentary - Too Big to Fail: Large Samples and the p-Value Problem." Information Systems Research (24:4), 2013, pp. 906-917.

[34] Dunteman, G. H. "Principal Component Analysis." 1989, Sage Publications.

[35] Wooldridge J. M. "Econometric Analysis of Cross Section and Panel Data." MIT Press, 2010, Cambridge, MA.

[36] Chapman, H. A., Kim, D. A., Susskind J. M. and Anderson, A. K. "In Bad Taste: Evidence for the Oral Origins of Moral Disgust." Science (323:5918), 2009, pp. 1222-1226.

[37] Jaspers, M. W. M., Steen, T., van der Bos, C. and Geenen, M. "The Think Aloud Method: a Guide to User Interface Design." International Journal of Medical Informatics 73, 2004, pp. 781-795.