An Experimental Evaluation of a De-biasing Intervention for Professional Software Developers

Martin Shepperd
Brunel University London
London, UK
martin.shepperd@brunel.ac.uk

Carolyn Mair
Fashion.Psychology
London, UK
carolyn.mair@gmail.com

Magne Jørgensen
Simula Research Laboratory
Oslo, Norway
magnej@simula.no

ABSTRACT

Context: The role of expert judgement is essential in our quest to improve software project planning and execution. However, its accuracy is dependent on many factors, not least the avoidance of judgement biases, such as the anchoring bias, arising from being influenced by initial information, even when it's misleading or irrelevant. This strong effect is widely documented.

Objective: We aimed to replicate this anchoring bias using professionals and, novel in a software engineering context, explore de-biasing interventions through increasing knowledge and awareness of judgement biases.

Method: We ran two series of experiments in company settings with a total of 410 software developers. Some developers took part in a workshop to heighten their awareness of a range of cognitive biases, including anchoring. Later, the anchoring bias was induced by presenting low or high productivity values, followed by the participants' estimates of their own project productivity. Our hypothesis was that the workshop would lead to reduced bias, i.e., work as a de-biasing intervention.

Results: The anchors had a large effect (robust Cohen's $d = 1.19$) in influencing estimates. This was substantially reduced in those participants who attended the workshop (robust Cohen's $d = 0.72$). The reduced bias related mainly to the high anchor. The de-biasing intervention also led to a threefold reduction in estimate variance.

Conclusion: The impact of anchors upon judgement was substantial. Learning about judgement biases does appear capable of mitigating, although not removing, the anchoring bias. The positive effect of de-biasing through learning about biases suggests that it has value.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; **Measurement**; **Experimentation**; **Estimation**; • **Social and professional topics** → *Project management techniques*;

KEYWORDS

Software engineering experimentation, Software effort estimation, Expert judgement, Cognitive bias

ACM Reference Format:
Martin Shepperd, Carolyn Mair, and Magne Jørgensen. 2018. An Experimental Evaluation of a De-biasing Intervention for Professional Software Developers. In *SAC 2018: SAC 2018: Symposium on Applied Computing , April 9–13, 2018, Pau, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3167132.3167293

SAC 2018, April 9–13, 2018, Pau, France

ACM ISBN 978-1-4503-5191-1/18/04…$15.00
https://doi.org/10.1145/3167132.3167293

1 INTRODUCTION

Effective management of software projects demands, amongst other things, accurate resource predictions. For this reason cost or effort modelling has been a major topic of research over many years [17, 26]. However, the preponderance of this research has focused on the development and evaluation of formal predictive models. In contrast, the role of human experts — who engage in this process, who make choices about model inputs and outputs — has been somewhat neglected [13, 14].

Human judgement and decision-making has been studied for decades by cognitive psychologists, e.g., the well known work of Kahneman et al. [21, 33]. An important finding is that humans typically use heuristics, i.e., simple mental strategies which, although sufficient in most circumstances, may lead to poor judgements and decisions in others and software engineering is not exempt from this.

When the use of heuristics leads to a deviation from a rational norm, such as when the heuristic does not fit the context or is based on misleading or irrelevant input, it leads to errors we call judgement and decision biases. Heuristics, and consequently the judgement and decision biases, are frequently unconscious. This means that the users of heuristics typically will not be able to explain properly how a judgement and or decision was made, why a poor judgement or decision was made or know how to improve the judgement and decision process.

Many judgement and decision biases have been identified, however our study focuses on the impact of the anchoring bias. This bias is thoroughly documented as widespread and leading to significant distortions of judgement [5, 22].

A judgement based on the anchoring heuristic, e.g., when estimating effort or productivity, may frequently be useful. Imagine a situation where a technically competent project leader indicates that she believes that a software development task should take about 10 work-hours. You are then asked about giving your judgement about the effort you would need for that task. Given that the project leader is competent, it saves you time and mental effort to base your thinking process on that 10 work-hours as a good

starting point, or to compare the current task with other tasks with size of about 10 work-hours to find out whether this is larger or smaller. It may even improve the accuracy of the effort estimate. But what if the number used by your anchoring heuristics is totally irrelevant, such as the number of hours spent on your previous task, or misleading, such as a very low number of work-hours a technically incompetent client wants you to use? Several studies suggest that software professionals, like everyone else, are affected by presented numbers, even when they are irrelevant or misleading [10, 15]. This happens even when professionals are explicitly requested to ignore them [18, 29].

While there are hundreds of studies on the presence of human biases in judgement and decision making, including many on the anchoring bias, there has not been much research on the impact of increased awareness of cognitive biases on the reduction of such biases (i.e., de-biasing). To investigate this topic, we conducted an experiment where the intervention was a workshop to increase participant awareness of cognitive biases and then compared these results with those of participants from a previously published experiment completing the same task who had not attended the workshop [16].

Another limitation of previous research is that most evidence for the existence of the anchoring bias comes from student samples and the use of tasks where students have little previous experience. In contrast, the sample in our study comprised professional software developers who were asked to estimate their own productivity on a task they had previously completed. This, we believe, makes the task more familiar for the subject and increases the relevance of the results to real-world tasks.

The remainder of the paper is organised as follows. First we present related work and supporting evidence for cognitive biases and how this might impact judgement and decision making. Next we describe the two-factor experimental design (low and high anchor, de-biasing and no intervention) experiment. We present the results of our robust statistical analysis, initially from 118 professional participants and then pooled with participants from a set of previous experiments. We conclude by discussing the implications of these results for improving professional judgements and outline some areas for further investigation.

2 RELATED WORK

The anchoring bias is one of the strongest, easiest to create, robust, long-lasting and studied of the human biases [9]. The most famous study of the anchoring bias involved a rigged wheel of fortune and the question: What percentage of the members of the UN are African countries? First, the research participants span the wheel, which stopped at 10 or 65 depending on how the wheel was rigged, and were asked whether they thought the percentage African countries in the UN was more than or less than the number on the wheel. Following that question, the participants were asked to predict the proportion of African countries in the UN. The difference in answers between the two groups was large. Those in the first group (wheel stopping at 10) gave a median prediction of 25% African countries in the UN, while those in the second group (wheel stopping at 65) gave a median prediction of 45% [33]. It is hard to imagine that the participants would think that a number on a wheel of fortune, which they believed gave a random number between 0 and 100, revealed any information about the actual proportion of African countries in the UN. Nevertheless, they were strongly affected by the number presented to them. Numerous subsequent studies, following similar anchoring inducing procedures, have shown similar effects. Even completely irrelevant anchors, such as digits from social security numbers or phone numbers, have been demonstrated to strongly bias people's judgements [2]

The anchoring bias is clearly relevant outside artificial experimental settings. Software professionals' time predictions were for example strongly affected by knowledge about what a customer had communicated as her expectation of time usage, in spite of being informed that the customer had no competence in predicting the time usage [18]. When asking these professionals whether they thought they had been affected by the customer's expectations, i.e., by the anchoring information, they either denied it or responded that they were just affected a little. This feeling of not being much affected, when in reality being affected a lot, is part of what makes the anchoring bias potent and hard to avoid. Even extreme anchors or suggestions, for instance that the length of a whale is 900 metres (unreasonably high anchor) or 0.2 metres (unreasonably low anchor), is effective in influencing people's judgements [32]. Anchoring effects seem to be pretty robust to all kinds of warnings. The following are instructions from a software development effort estimation study on anchoring: *I admit I have no experience with software projects, but I guess this will take about two months to finish. I may be wrong, of course; we'll wait for your calculations for a better estimate* [1]. In spite of the warnings, the software developers were strongly affected by the anchoring value of two months.

The cognitive basis of the anchoring bias is disputed and there are at least three different, partly overlapping, explanations: 1) Anchoring as communication (the attitude change theory), i.e., that it is natural for us to give weight to what other people communicate [34]. 2) Anchors as a starting point (the anchoring and adjustment theory), i.e., that the anchor is the starting point and that the adjustment away from the anchor typically is insufficient [21]. 3) Anchors as an activating experience (the selective accessibility theory), i.e., that the anchor activates experiences and that recently activated experience is more likely to be used in the subsequent judgement process [28]. All explanations have supporting evidence and it is possible that they all contribute to the observed anchoring bias.

De-biasing is applying mitigating interventions to reduce the impact of a bias. Fischhoff [8] suggests a fourfold classification scheme:

- (a) warning about the possibility of bias without specifying its nature.
- (b) describing the direction (and possibly extent) of the bias that might typically be observed.
- (c) providing feedback, preferably at a personal level.
- (d) offering an extended program of training with feedback, coaching, etc

Given the large effect size and importance of the anchor bias, it is not surprising that research has been devoted to study de-biasing strategies, including how to reduce or remove the anchoring bias. Although several methods for de-biasing have been proposed and tested, researchers have struggled to remove this effect. Examples

of de-biasing strategies with some positive effect, but far from eliminating the bias, are to "consider the opposite" [30] and introduction of new, more relevant, anchors (known as re-biasing) [23]. The study by Lovallo and Sibony [24] reported that the 25% companies best at avoiding and reducing decision biases, i.e., better at de-biasing, had a 5.3% advantage over the 25% worst (i.e., 6.9% vs 1.6% typical ROI). This suggests that de-biasing strategies are of substantial real-world importance.

In our paper, we examine the de-biasing effect of increasing the awareness of the anchoring effect among software developers. The evidence in support of this type of de-biasing is mixed. A positive, although not very large, effect of a training-based increase of bias awareness, including the anchoring bias, was reported in [27]. In contrast, no positive effect was found from teaching-based increase of bias awareness by [31]. The study reported by Welsh et al. [36] found a positive effect from increased bias awareness on the overconfidence bias, but none for the anchoring bias. The general finding seems to be that increased bias awareness typically has moderate to no effect on how much people are biased in their judgements and decisions [20]. No prior studies have, as far as we know, reported on the effect of increased anchoring bias awareness in the context of professional software developers.

3 EXPERIMENTAL METHOD

3.1 Participants

This study is based upon two series of experiments. The first was conducted by MJ and involved 292 participants from industry with no workshop (de-biasing) intervention. The second series were conducted by CM and MS with MJ involved for the first experiment of the second series. These experiments replicated the initial experimental design (this is documented in [16] as Estimation Task 1). In addition, the de-biasing intervention of a workshop was introduced prior to the actual experimental task. Table 1 shows the counts of participants by treatment. The participants were all professional software developers drawn from a total of 15 companies and seven different countries as indicated by Table 2. They were recruited as volunteers from companies with whom MJ had previously collaborated. This was supplemented by attendees from effort estimation workshops delivered by MS and CM in the UK and New Zealand.

Workshop?	High Anchor	Low Anchor	Total
N	142	150	292
Y	60	58	118
Total	203	210	410

Table 1: Participants by Treatment

3.2 Experimental Design

The participants were randomly allocated to either the high anchor or low anchor group. Each group was then given separate anchor values. The low anchor was based on the question "Do you believe your coding productivity was greater than 1 LOC per hour on your last project?". By contrast, the high anchor was based on the question "Do you believe your coding productivity was less than 200 LOC per hour on your last project?". Participants recorded the response, 'Yes' or 'No'. They were then all asked to report their estimate of programming productivity in LOC per hour. The actual estimates are used for this analysis.

Country	Count
Nepal	59
New Zealand	18
Poland	92
Romania	48
United Kingdom	16
Ukraine	114
Vietnam	63
Total	410

Table 2: Participants by Country

The de-biasing intervention comprised a 2–3 hour workshop on cognitive bias and estimation given immediately prior to the above task. Participants were introduced to the concept of cognitive bias and given examples from the psychology literature demonstrating the influence of bias on decision making. The biases covered included over-optimism and over-confidence [35], planning fallacy [4], peak-end rule [19], dual-process theory [6], blind spot bias [11] and anchoring [33]. The workshop concluded with a discussion on the influence of bias on prediction and estimating. In terms of Fischhoff's [8] classification scheme of de-biasing interventions we (b) described the direction and possible extent of the bias and (c) provided some personal feedback via an example task.

3.3 Data Collection and Cleaning

We recorded the following information from each participant summarised in Table 3.

Variable	Explanation
P_id	Unique participant id
Workshop	Y or N depending on the use of a de-biasing intervention
Block	Specific id of the experiment, e.g., there are multiple deliveries for some companies either at different times or locations.
Company	The employing company of the software developer - anonymised
Country	The country where the software development company is located
Anchor	High or low depending on the randomly allocated treatment
EstProd	Estimated coding productivity in LOC per hour for the last completed software project. This is the response variable.

Table 3: Data Collected

In terms of data cleaning we discarded participants who estimated their productivity as:

- missing values (5 cases eliminated)

- zero values as this implied that the participant had not engaged in coding (3 cases eliminated)
- excessively high values of ⩾ 500 LOC per hour since this implies an implausible level of productivity of almost one LOC per 7 seconds! (4 cases eliminated)

A representative sample of five rows of the data are given in Table 4. The raw data and R scripts are available from https://doi.org/10.6084/m9.figshare.5414200.v3 .

4 RESULTS

4.1 Summary statistics

In this section we present the results of our analysis of the experimental data. First we give the basic descriptive statistics for the response variable Estimated Productivity, then explore the basic anchoring effect and finally our main intervention: the de-biasing effect of the workshop.

Table 5 describes our response variable Estimated Productivity. We see values that range from 0.5 to 300 (after the data cleaning described in Section 3.3) with a strong positive skew (evidenced by the mean being greater than the median and the strong deviations particularly of the upper tail in the qqplot (Fig. 1). For this reason we also compute a 20% trimmed mean and standard deviation as more robust estimators [37]. Both are less than their untrimmed counterparts due to the positive skew (Table 5).

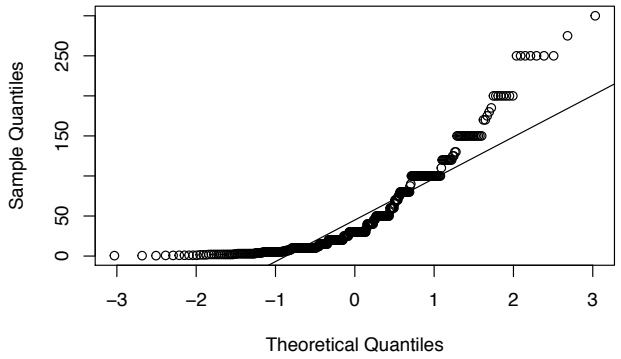

Figure 1: Estimated productivity qqplot

The qqplot also reveals the presence of many ties (horizontal segments of the curve) which correspond to popular round numbers. For example there are no predictions of 9LOC, but 41 of 10LOC and two of 11LOC. This is illustrated clearly by the stem and leaf plot where we see zero dominates as a trailing digit, followed by a five (see Fig. 2). Perhaps even more remarkable is that not one participant made an estimate ending in a nine. We conjecture that there is a high degree of uncertainty in the estimates which leads participants to use 5, 10, 20, ... rather than 9 (which would suggest a strong belief in estimation accuracy). For a discussion of the rounding phenomenon see [12].

```
  0 | 1111111111222222222222223333333333333333444444455555555555555555555555566666777778888
  1 | 00000000000000000000000000000000000000000112222555555555555555558
  2 | 000000000000000000000000000000003555555555558
  3 | 0000000000000000000000000000000000000055
  4 | 000000000000055568
  5 | 00000000000000000000000000005
  6 | 000000005
  7 | 000000555
  8 | 0000000000000000008
  9 | 0
 10 | 00000000000000000000000000000000000000000
 11 | 0
 12 | 0000000000055
 13 | 00
 14 |
 15 | 00000000000000000000
 16 |
 17 | 005
 18 | 05
 19 |
 20 | 00000000
 21 |
 22 |
 23 |
 24 |
 25 | 0000000
 26 |
 27 | 5
 28 |
 29 |
 30 | 0
```

Figure 2: Estimated productivity as a stem and leaf plot

4.2 The Anchor Effect

Recall that we do not know the true productivity levels for each participant. But given the random allocation of participants to the anchor treatments we do not believe there is any good reason to expect one group to be more productive than the other. The first thing to observe is the impact of the anchor on all participants, shown graphically in Fig. 3 as boxplots. Note the presence of extreme outliers, denoted by individual observations, for both anchor treatments. Note also the substantial difference in medians, shown by the line across each box and the 95% confidence limits for the medians shown by the notches which do not overlap.

More formally we can compare the two samples using the robust Yuen test with bootstrap to estimate the 95% confidence interval. The impact of the anchor is statistically significant, $p \approx 0$. The trimmed mean difference is 60.5 and the 95% confidence interval is (51.6, 69.4). In terms of effect size, this is either simply the trimmed mean difference of $\sim$ 60 LOC per hour between a low and high anchor estimate. Alternatively, if we want to standardise the effect size we can compute a robust version of Cohen's d using a pooled trimmed standard deviation which yields $\sim$ 1.18, an effect size which is between large and very large (0.8–1.3) [7]. Essentially when software professionals are asked to estimate coding productivity, the percentage difference between the low and high anchor groups was approximately 350%.

4.3 The Workshop Effect

So having shown that the anchor effect is very strong in the context of software estimation, we next consider the impact of the de-biasing intervention of the workshop. But first, we need to address a potential confounder in that the study design is unbalanced; we can see that there are experimental blocks that didn't receive the intervention at all, or vice versa (see Table 6). This is potentially problematic as the productivity estimates also differ considerably by country (see Table 7). The UK shows much lower Estimated Productivity and Nepal and Vietnam much higher than the other

P_id	Workshop	Block	Company	Country	Anchor	Est_Prod
P130	N	G	G	Ukraine	high	100.0
P334	Y	I3	I	Poland	high	20.0
P318	Y	I3	I	Poland	low	1.5
P250	N	K	K	Vietnam	low	15.0
P10	N	A	A	Romania	high	80.0

Table 4: Example Data Collected

Count	Mean	Median	SD	Min	Max	Trim mean	Trim SD
410	52.7	30	58.7	0.5	300	37.5	51.0

Table 5: Summary Statistics for Estimated Productivity

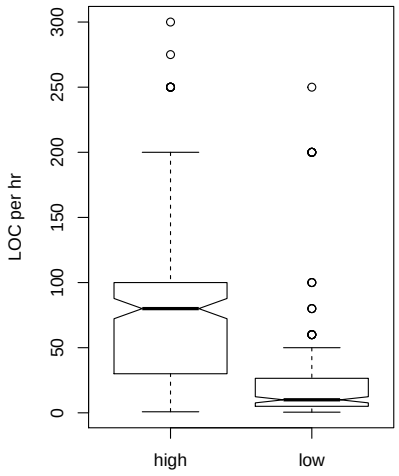

Figure 3: Boxplots of Estimated Productivity by Anchor Value

countries. Therefore we exclude the UK, Nepal and Vietnam to mitigate this problem. This leaves 272 participants with 102 receiving the de-biasing intervention.

We compare the estimates visually in Fig. 4 that groups participants both by anchor (low or high) and by de-biasing treatment (Y or N). It is clear from the boxplots that the median Estimated Productivity for the high anchor without de-biasing (high.N) is substantially greater than the median with de-biasing (high.N). Recall that the notches indicate the 95% confidence limits and note that these do not overlap. As indicated by the size of the whiskers, the spread of estimates also seems greater when there is no de-biasing workshop. Likewise, we see extreme outliers particularly for the no workshop condition. However, the effect for the low anchor is less obvious.

Country	N	Y
Nepal	59	0
NZ	0	18
Poland	47	45
Romania	48	0
UK	0	16
Ukraine	75	39
Vietnam	63	0

Table 6: Frequency Count of De-biasing Treatment by Country

Country	Mean EstProd
Nepal	75.39
NZ	33.4
Poland	41.0
Romania	51.4
UK	14.6
Ukraine	47.0
Vietnam	75.3

Table 7: Mean Estimated Productivity by Country

The median estimate of hourly productivity for the high anchor is reduced from 100 to 30 LOC/hr but for the low anchor the median remains unchanged at 10 LOC/hr. There are three possible reasons for the similarity of the median estimates for those in the low anchor group. First, the companies, and their software professionals, in the workshop group may have been more productive and as a consequence produced even lower estimates in a no workshop context. Second, it is harder to influence people to be negative about one's own performance, i.e., that there is less room for de-biasing interventions for the low anchor. Third, the de-biasing intervention may have increased their awareness of the optimism-inducing effect of anchor values, which in this case is the increase in productivity values through a high anchor, but not so much the optimism-reducing effect, corresponding to a low productivity anchor. More studies are needed to analyse and better understand this potentially interesting finding.

We also tabulate comparisons of means and, in parentheses, standard deviations in Table 8 and robust analogues based on 20% trimming in Table 9. Since trimming tends to remove extreme values we see the general effect is to slightly reduce our estimates of centre and dispersion.

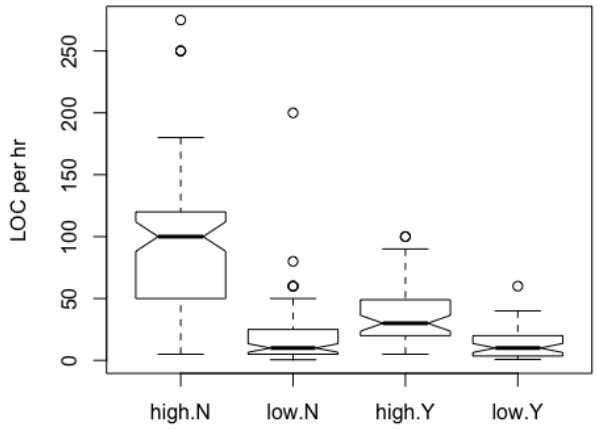

Figure 4: Boxplots of Estimated Productivity by De-biasing Intervention
Legend: high.N = high anchor, no de-biasing; low.N = low anchor, no de-biasing; high.Y = high anchor, de-biasing; low.Y = low anchor, de-biasing

Anchor	No workshop	Workshop
high	92.85	37.73
	(53.72)	(27.34)
low	19.17	13.20
	(25.89)	(12.67)

Table 8: Mean and Standard Deviations for Estimated Productivity by Anchor and De-biasing Workshop

Anchor	No workshop	Workshop
high	86.44	31.42
	(58.51)	(22.20)
low	13.08	10.18
	(14.05)	(10.26)

Table 9: 20% Trimmed Mean and Standard Deviations for Estimated Productivity by Anchor and De-biasing Workshop

Formally we can compare the central tendency and dispersion of the two conditions. For central tendency we apply the robust Yuen's test and find the trimmed mean difference is 25.6, $p \approx 0$ and the 95% confidence interval is (15.5, 35.7). This strongly suggests that the de-biasing workshop reduces estimates of productivity. Inasmuch as the higher estimates are influenced upwards by the anchor this is a desirable outcome.

However, we might also expect the spread of estimates to be narrowed if the effect of the anchors are reduced. To compare spread or dispersion we use a simple robust test to compare variance. We expect the de-biasing to reduce the variance of the estimates since the anchors will have less impact and not stretch out the distribution of estimates. Robust 20% trimmed estimates of standard deviation are given in Table 10 which indicates that the standard deviation is reduced about threefold with the de-biasing workshop intervention. As a formality we test that this reduction is significant. Since we already know the distribution is heavy-tailed, skewed and generally non-Gaussian, we use the Brown-Forsythe median variant of Levene's test of homogeneity of variance [3]. This gives a Test Statistic of 36.3, $p \approx 0$ meaning it is highly likely the two groups have different variances.

No workshop	Workshop
54.63	17.04

Table 10: 20% Trimmed Standard Deviations for Estimated Productivity by De-biasing Workshop

Considering both factors, the Anchor and the de-biasing Workshop together we use ANOVA, specifically the robust 2-way between-between method of Wilcox [25, 37]. The results are given in Table 11 however, we need to sound a note of caution. The variance is strongly heteroscedastic, the data imbalanced and therefore there may be ordering effects, so we only consider gross outcomes. There is strong evidence that both Anchor and Workshop are associated with estimated productivity, Anchor more so. It is also clear there is an interaction between Anchor and De-biasing confirmed by the Interaction Plot (Fig. 5). Essentially the de-biasing intervention only seems to impact the high anchor condition. This might be because (i) negative values for the estimate are meaningless and (ii) as we suspect the many of the higher values e.g., greater than 100 LOC/hr are somewhat hard to accept. Therefore it is probable that the high anchor is causing more bias or distortion than the low anchor.

Factor	F	p
Anchor	192.5	< 0.001
Workshop	72.2	< 0.001
Anchor:Workshop	58.4	< 0.001

Table 11: Robust 2-way Analysis of Variance

To summarise, we have strong evidence of both the anchor effect and a mitigating effect from the de-biasing workshop. In terms of effect size, this is either simply the trimmed mean difference of 26 LOC per hour between an estimate with and without de-biasing. (This is substantial but less than the Anchor effect). If we want to standardise we can compute a robust version of Cohen's d using a pooled trimmed standard deviation giving $d \sim 0.72$ which suggests a medium to large effect (0.5 – 0.8) [7]. Alternatively the impact of de-biasing can be assessed by considering the reduction in the spread of estimates (since the anchors will have a reducing distorting effect as a consequence of the de-biasing). We find that the standard deviation of the de-biased estimates is reduced about threefold so again support for the impact of our de-biasing workshops.

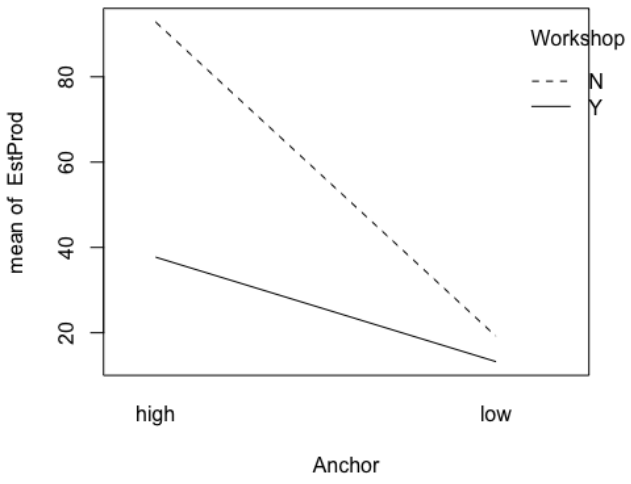

Figure 5: Interaction Plot of Anchor and De-biasing Intervention

5 DISCUSSION AND CONCLUSIONS

In this study we have addressed the real world problem of how biases, specifically the anchoring bias, influence software professionals making estimates and then how they might be mitigated. To do this we have conducted a series of experiments across seven countries with 410 participants. We believe this study is important because despite the emphasis on formal prediction systems, project cost decisions are ultimately made by humans, and these judgements are infrequent, but of high value. Therefore they cannot be conceived of as purely technical problems.

Our experiments yield four main findings.

(1) The effect of anchors on software professionals performing estimation tasks, in line with previous studies, such as [23], is very strong.
(2) The de-biasing workshop significantly reduces — but does not eliminate — this bias.
(3) The workshop also substantially reduces the variability in the estimates of professionals approximately threefold
(4) The workshop has a greater impact for the high rather than low anchor (although given the meaninglessness of a negative estimate, low estimates could only change in one direction).

However, there are some limitations to this work. First, we have only considered one type of bias and a relatively simple de-biasing intervention based on a 2-3 hour workshop. There are many other cognitive biases and judgement fallacies, at least some of which could be relevant to software engineering. Another limitation is that we don't know how long the de-biasing effect will last, but it is quite possible it is only transient. Therefore follow up work might be useful.

Nevertheless, this study has practical significance. It shows how professionals can be easily misled into making highly distorted judgements. This matters in that despite all our tools and automation, software engineering remains a profession that requires judgement and flair. Fortunately, we show, that it is possible to reduce, although not eliminate, these deleterious effects. There may well also be considerable scope for refining and improving de-biasing interventions.

ACKNOWLEDGEMENTS

This work was funded by EPSRC Grants EP/I038225/1 and EP/1037881/1. We are also grateful to the participants of the experiments.

REFERENCES

[1] J. Aranda and S. Easterbrook. 2005. Anchoring and Adjustment in Software Estimation. In *ESEC-FSE'05*. ACM Press, 346–355.
[2] D. Ariely, G. Loewenstein, and D. Prelec. 2003. "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics* 118, 1 (2003), 73–106.
[3] M. Brown and A. Forsythe. 1974. Robust tests for the equality of variances. *Journal of The American Statistical Association* 69, 346 (1974), 364–367.
[4] R. Buehler, D. Griffin, and M. Ross. 1994. Exploring the 'Planning Fallacy': why people underestimate their task completion times. *Journal of Personality & Social Psychology* 67, 3 (1994), 366–381.
[5] R. Buehler, J. Peetz, and D. Griffin. 2010. Finishing on time: When do predictions influence completion times? *Organizational Behavior and Human Decision Processes* 111, 1 (2010), 23–32.
[6] M. Deutsch and H. Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology* 51, 3 (1955), 629–636.
[7] P. Ellis. 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* Cambridge University Press.
[8] B. Fischoff. 1981. *Debiasing.* Technical Report. DECISION RESEARCH EUGENE OR.
[9] A. Furnham and H. Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics* 40, 1 (2011), 35–42.
[10] T. Halkjelsvik and M. Jørgensen. 2012. From origami to software development: A review of studies on judgment-based predictions of performance time. *Psychological Bulletin* 138, 2 (2012), 238–271.
[11] Katherine Hansen, Margaret Gerbasi, Alexander Todorov, Elliott Kruse, and Emily Pronin. 2014. People Claim Objectivity After Knowingly Using Biased Strategies. *Personality and Social Psychology Bulletin* 40, 6 (2014), 691–699.
[12] C. Jansen and M. Pollmann. 2001. On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics* 8, 3 (2001), 187–201.
[13] M. Jørgensen. 2004. A review of studies on expert estimation of software development effort. *Journal of Systems and Software* 70, 1 (2004), 37–60.
[14] M. Jørgensen. 2007. Forecasting of software development work effort: Evidence on expert judgement and formal models. *International Journal of Forecasting* 23, 3 (2007), 449–462.
[15] M. Jørgensen and S. Grimstad. 2011. The impact of irrelevant and misleading information on software development effort estimates: A randomized controlled field experiment. *IEEE Transactions on Software Engineering* 37, 5 (2011), 695–707.
[16] M. Jørgensen and S. Grimstad. 2012. Software Development Estimation Biases: The Role of Interdependence. *IEEE Transactions on Software Engineering* 38, 3 (2012), 677–693.
[17] M. Jørgensen and M. Shepperd. 2007. A Systematic Review of Software Development Cost Estimation Studies. *IEEE Transactions on Software Engineering* 33, 1 (2007), 33–53.
[18] M. Jørgensen and D. Sjøberg. 2004. The impact of customer expectation on software development effort estimates. *International Journal of Project ManagementÊ* 22, 4 (2004), 317–325.
[19] D. Kahneman. 1999. *Objective happiness.* Russell Sage Foundation, New York, 3–25.
[20] D. Kahneman, D. Lovallo, and O. Sibony. 2011. Before you make that big decision. *Harvard Business Review* 89, 6 (2011), 50–60.
[21] D. Kahneman, P. Slovic, and A. Tversky. 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press, Cambridge, UK.
[22] J. Klayman and K. Brown. 1993. Debias the environment instead of the judge: an alternative approach to reducing error in diagnostic (and other) judgment. *Cognition* 49, 1–2 (1993), 97–122.
[23] E. Løhre and M. Jørgensen. 2016. Numerical anchors and their strong effects on software development effort estimates. *Journal of Systems and Software* 116

(2016), 49–56.
[24] D. Lovallo and O. Sibony. 2010. The case for behavioral strategy. *JMcKinsey Quarterly* 2010, 2 (2010), 30–43.
[25] P. Mair and R. Wilcox. 2016. *Robust Statistical Methods in R: Using the WRS2 Package.* Technical Report. Harvard University. https://rdrr.io/rforge/WRS2/f/inst/doc/WRS2.pdf
[26] R. Malhotra. 2015. A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing* 27 (2015), 504–518.
[27] C. Morewedge, H. Yoon, I. Scopelliti, C. Symborski, J. Korris, and K. Kassam. 2015. Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences* 2, 1 (2015), 129–140.
[28] T. Mussweiler and F. Strack. 1999. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology* 35, 2 (1999), 136–164.
[29] T. Mussweiler and F. Strack. 2001. The Semantics of Anchoring. *Organizational Behavior and Human Decision Processes* 86, 2 (2001), 234–255.
[30] T. Mussweiler, F. Strack, and T. Pfeiffer. 2000. Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin* 26, 9 (2000), 1142–1150.
[31] G. Oliver, G. Oliver, and R. Body. 2017. BET 2: Poor evidence on whether teaching cognitive debiasing, or cognitive forcing strategies, lead to a reduction in errors attributable to cognition in emergency medicine students or doctors. *Emergency Medicine Journal* 34, 8 (2017), 553–554.
[32] F. Strack and T. Mussweiler. 1997. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology* 73, 3 (1997), 437–446.
[33] A. Tversky and D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.
[34] D. Wegener, R. Petty, B. Detweiler-Bedell, and W. Jarvis. 2001. Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology* 37, 1 (2001), 62–69.
[35] N. Weinstein. 1980. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* 39, 5 (1980), 806–820.
[36] M Welsh, S Begg, and R Bratvold. 2007. Efficacy of bias awareness in debiasing oil and gas judgments. In *29th Annual Cognitive Science Society.* Cognitive Science Society, 1647–1652.
[37] R. Wilcox. 2012. *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). Academic Press.