International Journal of Forecasting 32 (2016) 112–120

Contents lists available at ScienceDirect

International Journal of Forecasting

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

A parsimonious explanation of observed biases when forecasting one's own performance

Sheik Meeran *, Paul Goodwin, Baris Yalabik

University of Bath, UK

ARTICLE INFO

Keywords:
Judgmental forecasting
Metacognitive skills
Regression effects
Self-performance forecasting
Anchoring and adjustment

ABSTRACT

Forecasting one's own performance on tasks is important in a wide range of contexts. Over-forecasting can lead to an unresponsiveness to advice and feedback. In group forecasting, under-forecasting may lead individuals to discount valuable inputs that they could contribute. Research shows that those who perform relatively poorly in tasks tend to make predictions that are too high, while high performers tend to under-forecast their performances. Several explanations have been put forward for this 'regressive forecasting', such as a lack of metacognitive skills in poor performers and a false-consensus bias in high performers. Others claim that the bias is simply an artefact of regression. In this study, people were asked to forecast their performances on six multiple-choice tests. The results suggest that a simple explanation based on the anchoring and adjustment heuristic would account for the phenomenon, at least in part.

1. Introduction

Being able to forecast one's future performance based on an accurate perception of one's abilities and skills can be important in a number of contexts. These include career choices, making personal assessments of one's need for education and training, and decisions where personal failures may lead to danger or extensive losses. Forecasting one's performances on different tasks can also be important to those involved in judgmental forecasting itself. For example, in sales forecasting, a tendency to over-forecast one's future performance might lead to a lack of responsiveness to advice and feedback (e.g., Bonaccio & Dalal, 2006; Dunning, 2013; Lim & O'Connor, 1995). Similarly, in group forecasting situations, such as applications of the Delphi method, a propensity to over-predict one's forecasting performance might lead one to overweight one's own forecasts relative to those of the group. This may reduce a panel member's willingness to change their judgment when they receive information on the forecasts of other group members. In contrast, under-forecasting one's performance or underestimating one's expertise may lead one to discount the potentially valuable inputs that one could add to the forecasting process (though Rowe & Wright, 1999, found the evidence linking confidence in one's forecast and a willingness to change to be inconsistent). These potential problems would apply in particular to group-based forecasting methods that require group members to self-rate their expertise explicitly (e.g., DeGroot, 1974).

A number of studies have investigated how accurate people are at forecasting their own performances on tasks, tests or examinations (Burson, Larrick, & Kayman, 2006; Clayson, 2005; Kennedy, Lawton, & Plumlee, 2002; Krueger & Mueller, 2002; Kruger & Dunning, 1999; Miller & Geraci, 2011). The results have varied from findings of no correlation between predicted and actual performances to findings of a significant correlation. However, even when there is a significant positive correlation, a common finding is that, on average, relatively poor performers tend to over-forecast their performances, while high performers tend to

* Correspondence to: School of Management, University of Bath, Bath, BA2 7AY, UK. Tel.: +44 1225 383954.
E-mail address: s.meeran@bath.ac.uk (S. Meeran).

http://dx.doi.org/10.1016/j.ijforecast.2015.05.001
0169-2070/

under-forecast how well they will do. Several explanations have been put forward for this phenomenon, which we will term regressive forecasting. For example, Kruger and Dunning (1999) have argued that poor performers are unaware of their own incompetence, while high performers suffer from a false consensus effect, in that they assume that their abilities are shared by their peers. Others have suggested that the bias is merely an artefact of regression (Krueger & Mueller, 2002). In this paper, we adopt a judgmental forecasting perspective in order to suggest an alternative explanation for this tendency, and test it empirically. Our explanation is more parsimonious than many others that have been suggested, and hence is consistent with Occam's razor, which states that the simplest hypothesis, involving the fewest assumptions, should be favoured (see for example Domingos, 1999, for a discussion of Occam's razor).

We begin by reviewing the literature relating to this topic, before developing a theoretical model to represent the forecasting process. We then present an analysis of data from six in-course multiple-choice tests of statistical and forecasting knowledge. This enables us to model the process used by people to forecast their own performances under conditions in which the outcome was important and consequential to the individuals involved. The consequences arose because the final grade of the students' degrees, or whether they were able to progress to the later stages of the course, depended partly on their performances in these tests. A key advantage of the use of multiple-choice tests in this research is that the scores achieved are determined objectively. The use of marks for an essay-based examination, for example, would introduce an additional element of variation, namely the subjective marking of the examiner. Thus, forecasting one's performance would be confounded with forecasting the subjective (and probably inconsistent) scoring of the marker. Of course, the choice and nature of the questions on a multiple-choice test is based on the subjective judgment of the examiner, but the extent of the contribution of this subjectivity to the student's mark is far less than in many other forms of performance assessment.

2. Literature review

Studies of individuals' abilities to forecast their performances on tests and examinations have considered forecasts of two types: (i) predictions of marks, scores or grades (e.g., Clayson, 2005; Kennedy et al., 2002; Miller & Geraci, 2011); and (ii) predictions of the percentile in which the mark score or grade would lie (e.g., Burson et al., 2006; Krueger & Mueller, 2002; Kruger & Dunning, 1999). A common finding has been that, while the low performers have produced forecasts that are too high, the high performers have tended to under-forecast their scores or percentile positions (Kennedy et al., 2002; Kruger & Dunning, 1999). Similar patterns of the unskilled overestimating and the skilled underestimating their skill levels have also been recorded in domains such as driving (Kunkel, 1971), reading (Maki, Jonas, & Kallod, 1994), and social skills (Fagot & O'Brien, 1994). A third finding has been that these errors in forecasts are asymmetric, in that they tend to be greater for the low performers (Krueger & Mueller, 2002; Kruger & Dunning, 1999). The reasons underlying these findings have generated much controversy, with a range of alternative explanations being put forward.

There are a number of factors that may lead to individuals over-forecasting their test performances. One well-known phenomenon is the "above average effect", where most people perceive their skills to be above average. This has been observed in areas ranging from football (Felson, 1981) to business management and leadership (Larwood & Whittaker, 1977). While the effect is associated with statistically illogical judgments, Krueger and Mueller (2002) argue that, from an individual perspective, such optimism can be rational. For example, optimism can also be a valuable source of motivation. However, this does not explain directly why the forecasts are regressive, in that the optimism is only associated with low performances. Nor does it explain directly the observed asymmetry in the errors. One partial explanation is that test scores and percentiles are bounded (e.g., between 0 and 100), so that the higher one's actual score is, the less scope there is for over-forecasting it. A more elaborate explanation is provided by Kruger and Dunning (2002), who argue that unskilled individuals lack metacognitive skills. Their lack of skill in a particular domain is associated with a lack of skill in assessing their ability in that domain. If a person is incompetent, they also lack the ability to realise their incompetence (see also Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). More recently, Dunning (2013) has presented several examples of unskilled individuals being unaware of their lack of skill, and mentions various implications for organisations, such as the difficulties of recognising expertise in groups (Bonner, Baumann, & Dalal, 2002; Cone & Dunning, 2011), a reluctance to seek advice (Bonaccio & Dalal, 2006), and the evaluation of feedback for the purpose of self-improvement (Mobius, Niederle, Niehaus, & Rosenblat, 2011; Sheldon, Ames, & Dunning, 2011). However, Kruger and Dunning's (2002) metacognitive hypothesis does not explain why high achievers often produce forecasts of their *percentiles* that are too low. For this, they turn to the false consensus effect (Ross, Greene, & House, 1977), and argue that the high achievers assume their peers to be as skilled as they themselves. Hence, they tend to assess their ability as being closer to the middle of the range of performances, when their true percentiles are higher. However, it is not exactly clear why this bias should be peculiar to high achievers. Also, it would not explain any tendency of high achievers to under-forecast their scores, as distinct from their percentiles. Although Kruger and Dunning did find that high achievers forecasted their scores reasonably accurately, other studies have found this bias to occur with forecasts of scores as well (Clayson, 2005; Miller & Geraci, 2011).

Later studies have reconsidered the notion that poor performers are unaware of their limitations. Of course, the term 'limitations' in this context can refer to a number of separate abilities. These include: (i) a lack of ability relating to the skill that is being assessed in the test; (ii) a lack of ability in the self-assessment of one's skill in this domain; (iii) a lack of ability to convert such a self-assessment into a forecast of one's score; and (iv) a lack of ability to appraise the probable accuracy of this forecast. Miller and

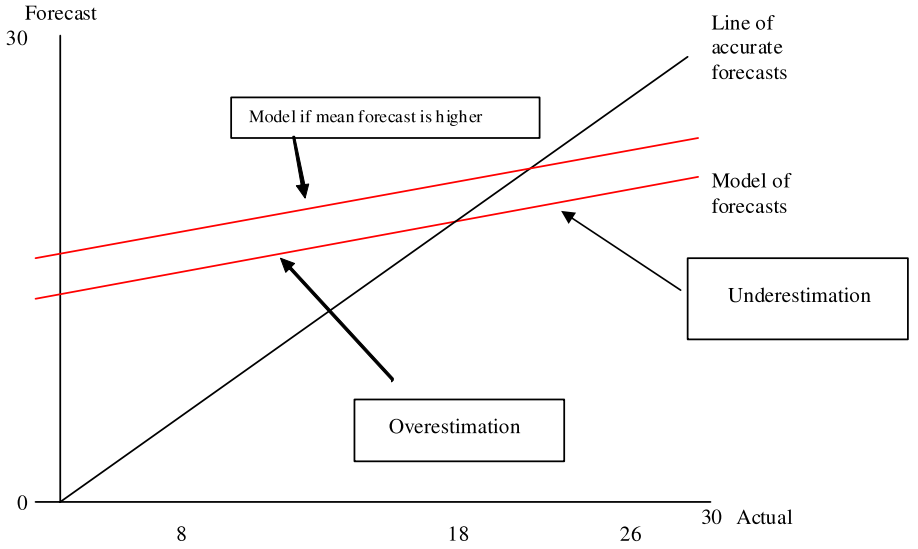

Fig. 1. The effect of increasing the mean forecast.

Geraci (2011) addressed (iv) and found poor performers to exhibit more uncertainty about their forecasts. They found that, while poor performers tended to over-forecast their grades, they also attached lower confidence ratings to their forecasts, suggesting that they are very much aware of their inability to assess their skill level accurately. One possible reason for these inaccurate assessments is suggested by Krajc and Ortmann (2008). They point out that most of the studies have been conducted in elite educational institutions, where there is a bunching of grades, and argue that this means that poor performers have a harder job than high performers in making inferences about their abilities — a so-called "signal extraction" explanation. However, Schlösser, Dunning, Johnson, and Kruger (2013) find no evidence that this is the explanation of the Kruger–Dunning effect.

In a direct critique of Kruger and Dunning's (1999) theory, Krueger and Mueller (2002) argue that the phenomenon is simply a result of the above average effect and a regression analysis. When the forecasts and actual scores on a test are correlated imperfectly and the variance of the forecast is less than the variance of the actual scores, the slope, b, of the regression line (Forecast $=$ $a + b$ Actual Score) will be less than one. Given that perfectly accurate forecasts would be represented by a line where $a = 0$ and $b = 1$, the two lines will cross, so that, on average, poor performers tend to over-forecast and high performers to under-forecast. (Alternatively, if the lines do not cross within the bounds of the scores determined by the test, the tendency to over-forecast will decline as an individual's actual performance improves.) Because of the above average effect, the intercept of the line, a, will tend to be relatively high (see Fig. 1), so that the average level of under-forecasting will be less than the average level of over-forecasting.

Note that regression factors will only account for the phenomenon when conditions lead to the line's slope being less than one. When the forecasts vary more than the actual scores (a situation that is entirely possible in judgmental forecasting, see O'Connor, Remus, & Griggs, 1993), this may not be the case. For example, if the scores on a test provide little discrimination between the worst and best performers, their variation may be much smaller than those of the forecasts, and the slope of the line could exceed one, depending on the correlation between the forecasts and scores. Thus, if regression is used to explain the phenomenon, it simply raises a new question as to why the slope of the line should be expected to be less than one.

Burson et al. (2006) also argued that metacognitive factors do not lead to the bias. They found that, when people were asked to predict the percentiles for their performances on a series of tests, the best and worst performers performed very similarly in their predictions of their performances when the task was of moderate difficulty. Moreover, the best performers actually produced less accurate forecasts than the worst performers on very difficult tasks. Burson et al. (2006) suggested that this inaccuracy in the percentile forecasts was due to a combination of noise (for example, the tests involve an element of luck, depending on which questions appear), biases (such as a tendency to anchor on their perception of their own performance), and the usefulness and availability of performance feedback.

Making a prediction of one's mark on a test is essentially a judgmental forecasting task. Judgment is used to integrate the perceived effects of the available cues (which may include previous test marks, a perceived 'norm' mark on the test, or one's perceived level of effort in preparing for the test) in order to forecast the value of an uncertain quantity, given that this value will not be known until the future (Armstrong, 2001, p. 790). In some aspects, the task parallels that of a sales person forecasting the sales that they will achieve next month. Here, they may use cues like the previous month's sales, a perceived 'normal' level of monthly sales, and their perceived effort in trying to generate sales. Despite this, few papers have referred to the judgmental forecasting literature when examining the tendency for high performers to under-forecast and poor performers to over-forecast their performances. This literature can potentially yield a number of new insights into the causes of

this bias and ways in which it can be measured. For example, forecasting research provides a set of potentially useful tools for analysing the relationship between forecasts and actual outcomes, and hence, for assessing the components of skill of the judgmental forecaster (Stewart & Lusk, 1994). High correlations between forecasts and outcomes are often quoted, but they can be misinterpreted (Kruger & Dunning, 2002) because they provide no information on systematic biases in forecasts. For example, the forecasts 1, 2, 3, 4 and 5 are perfectly correlated with the outcomes 21, 22, 23, 24 and 25, respectively, despite the fact that each forecast systematically underestimates the outcome by 20 units. If the outcomes were 2, 3, 4, 5 and 6, respectively, we would obtain the same correlation, even though the systematic underestimation is much less.

Judgmental forecasting researchers have also investigated the *processes* by which people arrive at their forecasts, typically by modelling their use of the available information (e.g., Lawrence & O'Connor, 1992). One common finding is that people often anchor on an initial estimate or forecast and then adjust from this, based on other information, in order to reach their final forecast (Tversky & Kahneman, 1974). For example, in time series extrapolation, in the absence of a trend, they tend to anchor on the most recent value and adjust from this to account for the mean of the series (Bolger & Harvey, 1993; Lawrence & O'Connor, 1992). A consequence of using this anchoring-and-adjustment heuristic is that the adjustment from the anchor is usually not sufficient to reflect the implications of the new information. The anchor can be particularly powerful when it is self-generated (Cervone & Peake, 1986; Strack & Mussweiler, 1997).

At least one study has hinted that people may use anchoring and adjustment when forecasting their test marks. Clayson (2005) states that his result "…suggests an interesting hypothesis. The students appear to be estimating their grades roughly half way from their actual scores to some norm. In this study, that norm appears to be the average GPA of the university". As has already been mentioned, Burson et al. (2006) also considered the possibility of anchoring. We therefore hypothesise that, when asked to forecast their test mark, people follow a process where they first assess a 'norm' mark, then adjust from this based on an assessment of their own ability as an individual. The assessment of the norm acts as an anchor, and the resulting forecast is a weighted average that lies *between* the two assessments. Thus, even if people can assess their own ability perfectly, the regressive forecasts that have been encountered in earlier studies would still be observed.

3. A theoretical model

Anderson's (1965) integration model suggests that anchoring and adjustment can be modelled as a weighted average of a starting or initial value (i.e., an anchor), G, and the estimate, U, that the person would have made had they not seen the anchor (e.g., see Choplin & Tawney, 2010). That is:

$$\text{Estimate} = (1 - w)\, G + wU + k, \tag{1}$$

where w is a weight and k is noise.

In our case, we assume that, if an individual is *free from anchoring*, their forecast of their mark would have the following linear relationship with their actual performance:

$$U = a + bA + i, \tag{2}$$

where A is the true mark, i is noise, and a and b are bias parameters that reflect the fact that students may have estimated factors such as the difficulty of the test wrongly. If $b = 0$, this would imply that there was no association between the student's unanchored forecast and their actual performance.

If the student is influenced by an anchor, their forecast (F) will be:

$$F = (1 - w)\, G + w[a + bA + i] + j, \tag{3a}$$

where j is noise, representing different susceptibilities to the anchor.

Thus:

$$\begin{aligned} F &= (1 - w)G + w[a + bA] + e \\ &= (1 - w)G + wa + wbA + e, \end{aligned} \tag{3b}$$

where $e = wi + j$.

We can also represent the relationship between the forecast and the actual mark as:

$$F = \alpha + \beta A + e. \tag{4}$$

So, from Eqs. (3b) and (4):

$$\alpha = (1 - w)G + wa, \quad \text{so } G = (\alpha - wa)/(1 - w),$$

and also $\beta = wb$, so $w = \beta/b$;

$$\text{hence, } G = \frac{1}{1 - w}\left[\alpha - \frac{\beta}{b}a\right]. \tag{5}$$

Remember that a and b are not observed directly. However, if the forecasts without the influence of an anchor are unbiased, $a = 0$, $b = 1$ and $w = \beta$, so G simplifies to:

$$G = \frac{\alpha}{1 - \beta}, \tag{6}$$

where α and β are estimated from the results of a cohort of students using regression analysis. If the participants in a given cohort were using the anchoring and adjustment heuristic (which we test for later), Eq. (5) or (6) yields an estimate of the mean value of the anchors that they used.

4. Data collection

To investigate the anchor-and-adjust hypothesis, we gathered data from six multiple-choice tests across three cohorts of students. In all cases, we focused on forecasts of test scores, rather than percentiles. Table 1 gives details of the cohorts, tests, and numbers of students involved. In the case of cohort 2, the tests (Tests 2, 3, 4 and 5) took place consecutively at roughly two-week intervals over the autumn (fall) semester.

The participants were three cohorts of undergraduate students at the University of Bath who were taking courses in business forecasting or business statistics. Although the forecasting course explored the biases associated with judgmental forecasting, the tests took place before this

Table 1
Details of the data sets.

Test	n	Subject	Cohort	Max possible mark	Mark forecast		Forecast of cohort mean				Actual mark	
					Mean	SD	Mean	Median	Range	SD	Mean	SD
1	96	Fcasting	1	30	20.8	4.0	n/a	n/a	n/a	n/a	17.2	3.7
2	118*	Fcasting	2	20	14.9	2.1	14.8	15	15	2.1	12.8	2.8
3	121	Fcasting	2	20	14.0	2.5	14.4	15	11	1.9	12.8	2.7
4	126	Fcasting	2	20	14.1	2.4	14.3	15	8	1.6	15.5	2.2
5	121*	Fcasting	2	20	14.0	2.4	14.4	15	8	1.6	13.0	2.1
6	82	Statistics	3	12	7.6	1.5	8.5	8	11	1.0	8.8	2.1

* One student was removed from each set. Their forecasts of the cohort mean mark were 1 and 2 out of 20, respectively.

Table 2
Estimates of the mean of the anchors used by participants.

Test	Estimate of alpha	Estimate of beta	R-squared	Estimated mean anchor	Mean forecast of cohort mean	Mean forecast of cohort mean as % of max mark
1	15.30	0.32	8.6%	22.5	n/a	n/a
2	14.00	0.14	2.3%	14.9*	14.8	74.1
3	9.88	0.32	11.9%	14.6	14.4	71.9
4	9.12	0.32	8.7%	13.5	14.3	71.4
5	9.09	0.38	10.2%	14.6	14.4	72.0
6	5.61	0.22	10.2%	7.2	8.5	70.8

Tests 2 and 6 were the students' first experience of tests of this type.
* The beta estimate was not significant, so the mean forecast is the estimated anchor.

topic was presented. No credit was given for participation. The students took multiple-choice tests, ranging in time from 20 to 45 min, that were designed to assess their understanding of the concepts that had been covered by the course prior to the test. Before the test, they were asked to forecast the mark that they expected to achieve. For Tests 2 to 6, they were then also asked to forecast the mean mark that they expected the cohort to achieve. They were assured that their forecasts would have no influence on their mark, and would be treated as confidential.

5. Analysis

To discover whether there was a typical anchor that the students used, the model in Eq. (4) was fitted to the data sets using a least squares regression. The results are shown in Table 2, which also shows an estimate, obtained from Eq. (6), of the mean of the anchors used by the participants, under the assumption that the unanchored forecasts were unbiased (an assumption that will be examined in a later section).

It can be seen from Table 2 that many of the parameter estimates for the different tests are very similar. In all cases, students scoring a mark below $\alpha/(1-\beta)$ would typically over-forecast their performance, while those scoring above this would typically under-forecast, which is consistent with earlier findings. However, it is particularly noteworthy that the estimates of the mean of the anchors used by participants are remarkably close to the average student forecasts of the cohort mean mark. The largest differences appear for Tests 2 and 6, where the students had no experience in taking these tests; however, even these differences are small. Also, the average forecasts of the cohort mean mark are found to be within the range of 70.8% to 74.14% of the maximum mark in all cases. These results are consistent with the students, on average, viewing a typical mark on the test as being about 72% of the maximum, and then adjusting from this, based on an assessment of their own ability, to obtain forecasts of their individual marks. We have not yet established that the unanchored forecasts were unbiased, but this demonstrates the possibility that students may have unbiased expectations of their performances and yet still produce biased forecasts, because of the effect of anchoring. In this case, regressive forecasts would still be possible without the need for more elaborate explanations.

The results above provide a prima facie case that, on average, the students were regarding a mark of around 72% as a 'norm' or typical mark, and then adjusting from this to take into account their own perceived ability. The estimates of the cohort mean mark appear to coincide with this perceived 'norm'. In this case, the R^2 values for the models in Table 2 will be low because the individual variation in the perceived norm (or expected cohort mean mark) is not taken into account. Before each of Tests 2 to 6, the participants were asked to provide a prediction of what the cohort mean would be. This enabled the forecasts of their individual mark (F) to be regressed on both their prediction of the cohort mean (G) and their actual mark (A), yielding models of the form:

$$F = \beta_0 + \beta_1 G + \beta_2 A + e. \tag{7}$$

Comparing this with Eq. (3b):

$$F = wa + (1-w)G + wbA + e,$$

it can be seen that:

$$\beta_0 = wa, \quad \text{so } a = \beta_0/w$$
$$\beta_1 = (1-w), \quad \text{so } w = 1-\beta_1$$
$$\beta_2 = wb, \quad \text{so } b = \beta_2/w.$$

Table 3 presents details of the models for Tests 2 to 6, and the resulting estimates of w, a and b. In all cases,

Table 3
Detecting biases in the unanchored forecasts.

Test	Estimate of β_0	Estimate of β_1	Estimate of β_2	R^2	Estimate of w	Estimate of a	Estimate of b
2	5.60***	0.54*	0.16***	30.5%	0.46	12.17	0.35
3	3.40	0.71***	0.14*	37.7%	0.29	19.31	0.48
4	1.49	0.58***	0.29***	23.6%	0.42	13.33	0.69
5	0.66	0.65***	0.30***	28.4%	0.35	16.00	0.86
6	2.66	0.34*	0.23**	15.7%	0.66	8.48	0.35

* Significant at the 5% level.
** Significant at the 1% level.
*** Significant at the 0.1% level.

the coefficient for the individual's prediction of the cohort mean mark is significant at the 5% level at least. Note that if $\beta_0 = 0$ and $\beta_1 + \beta_2 = 1$, this implies that the unanchored forecasts are unbiased (i.e., $a = 0$ and $b = 1$, deduced from $1 - w + wb = 1$).

Restricted least squares (e.g. Gujarati, 1995) was used to test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 + \beta_2 = 1$. In the case of Tests 2, 4 and 6, the hypothesis could be rejected with p-values of less than 0.001, 0.018 and less than 0.001, respectively, suggesting that the unanchored forecasts were biased. However, there was no evidence of bias in the unanchored forecasts for Tests 3 and 5 (all p-values were at least 0.942). Thus, for these two tests, the forecasts can be represented as a weighted average of the predicted cohort mean and an unbiased forecast of the mark.

Note that Tests 2 and 6 were each the students' first encounter with tests of this nature, and the regressive bias may be a result of a number of factors. These may include those already suggested in the literature, such as the 'unskilled and unaware' and 'false consensus' explanations, or may simply reflect the students' inability to produce accurate predictions of their marks. Because the tests were novel, the students would have little information on which to base their forecasts, and tests that were harder or easier than expected could create such a regressive effect. In particular, the results for Test 2 suggest that a major factor was the students' tendency to underestimate the difficulty of the test. Only students scoring above $a/(1 - b)$ marks would typically produce unanchored forecasts that underestimated their marks. This would be students scoring more than 19 marks out of 20, and thus, on average, almost all of the students would be expected to overforecast their marks. Test 4 had higher mean marks than the other tests, and may therefore have been easier than the students expected. However, even in the case of these three tests, the results suggest a further biasing effect as a result of anchoring.

In summary, our models provided evidence of anchoring in *every* case that we investigated. Table 3 shows that β_1, the coefficient for the anchor, was significant at the 5% level or less in every equation. It is important to draw a distinction between a biased anchor (which we did find in some cases) and having no anchor at all. If one is given an anchor of 1000 degrees Celsius when forecasting tomorrow's midday temperature in Seattle, the anchor is clearly biased; however, the individual may still be using the anchor and adjustment heuristic when making their forecast. Thus, the existence of a biased anchor does not mean that anchoring and adjustment was not being employed. Even when other theories that have been suggested in the literature may apply, such as the metacognitive explanation, anchoring may lead to additional biasing effects.

6. Discussion

The analysis above raises three questions:

1. Is it possible that other anchors were being used, rather than the prediction of a 'norm' mark, which appears to be represented by the expected cohort mean?
2. Is it possible that the prediction of the cohort mean anchored on the individual's forecast of their marks, rather than the other way round?
3. Why would values distributed around 72% act as anchors?

In this section we will begin by addressing these issues, before discussing various design issues associated with this study, such as why some tools and methods, such as verbal protocol analysis, were not used; why we used a task structure based on forecasting marks on multiple-choice tests; whether the framing of the questions would have had an impact on the quality of the data collected; and whether this task structure is suitable for drawing inferences in a teamwork setting.

To address the first question, the following alternative potential anchors were considered: (i) specific points on the marks scale, such as the mark achievable by guesswork or the midpoint of the scale; (ii) the student's previous test mark, if this existed; (iii) the student's mean mark on the previous two tests, if applicable; (iv) the student's mean mark on all previous tests, if applicable; and (v) the student's prediction of the cohort mean for the previous test, if applicable.

If other points on the mark scale had acted as a common anchor for the students, then we would expect the estimated mean of the anchors used by participants to have been equal to this value. As Table 2 indicates, this coincided with values of around 72% of the maximum marks, suggesting that none of these other points acted as a common anchor.

Fitting models of the form shown in Eq. (7), with G representing suggestions (ii) to (v) above, in turn, always led to R^2 values that were much lower than the values (shown in Table 3) where G equalled the predicted cohort mean (the R^2 values ranged from 9.0% to 16.1%, depending on the model and the test). Thus, there was no support at all for the possibility of alternative anchors.

The second possibility was that the predictions of the cohort mean did not act as the anchor; instead, they were

themselves anchored on the individual test mark forecasts. After all, in Tests 2 to 6, the students were asked for their forecast of the cohort mean directly after making a forecast of their own individual mark. Tests for the direction of causality in the regression models based on coefficients of the kurtosis were inconclusive (Pornprasertmanit & Little, 2012). However, there is some evidence that the cohort mean is the more likely anchor. First, in Test 1, the students were not asked to forecast the cohort mean before taking the test, but the estimated mean of the anchors they used, shown in Table 2, is very similar to those on the other tests (i.e., 75% of the maximum mark). Indeed, all of the other models referred to in Table 2 relate only to the individual mark forecasts, suggesting that these were anchored on the predicted cohort means *before* these cohort mean predictions were elicited formally.[1] In addition, Kruger (1999) predicts that the 'above-average' effect will prevail in situations in which ability levels are high, such as here, because 'people anchor on their assessment of their own abilities and insufficiently adjust to take into account the skills of the comparison group'. Thus, Kruger argues that estimates of individual performances form the anchor. However, Table 1 shows that, in four of the five tests for which information is available, individuals' mean forecasts of their own marks in our tests were below their forecasts of the cohort mean, meaning that we had a 'below-average' effect. This should not have occurred if Kruger's theory was applicable in these tests and individual performance was the anchor.

Finally, why would marks distributed around 72% to 75% act as anchors? The consistency of this across the tests was extraordinary, considering that Tests 2 to 5 were conducted over a period of eight weeks, while Tests 1 and 2 involved different cohorts and Test 6 involved a different subject. One possibility is that the students simply started their forecasts with a point midway between the 50% mark and the maximum mark (i.e., the 75% mark). There is some evidence that users' ratings of products on the internet tend to peak at around 70% of the maximum rating (e.g. Duan, Gu, & Whinston, 2008; Poundstone, 2014), so there may be a natural tendency to use values of around 70% as the starting value for estimation and forecasting. Also, when asked to choose a number on a scale from zero to nine, people most often choose seven (Kubovy & Psotka, 1976). Indeed, the predictions of the cohort mean may have been anchored themselves on values in the 70% to 75% region, before acting as an anchor for the individual marks forecast. It is possible to gain a mark of 100% on a multiple-choice test, whereas marks above 70% on an essay would be rare, so a starting value in this region may be seen to be feasible. It is also worth noting that there is a commonality between these cohorts, in that the entry requirements to their degrees are the same, requiring A grades (which is equivalent to 70% and above) in their pre-University examinations (such as GCE A-level examinations in the UK), so marks above 70% may be regarded as a norm.

Our use of regression models in this work has meant that our only way of infering whether anchoring and adjustment was typically being used was by modelling the average responses of a large sample of participants. Verbal protocol analysis would have offered an alternative approach (Epley & Gilovich, 2001); however, we note that its application in our work would have a number of limitations. The use of heuristics is an intuitive process. By requiring an explicit account of the judgmental process from respondents, a 'more conscious' process might come into play, so that the reported process might not reflect that which would have been used naturally. Moreover, the act of providing a verbal account of one's judgment process will itself use cognitive resources, thus reducing those available for the judgment task and potentially distorting the process that would otherwise have been used. Protocols are also difficult to analyse formally, and such difficulties may lead to unreliable inferences (Russo, 1978). In addition, the approach normally only allows small samples to be used, due to its high demand for resources. Indeed, Carroll and Johnson (1990) argue that "sometimes it is clearly not worth the effort to do protocol analysis.... [In some cases] developing some form of weighted-average model is more likely to be cost-effective" (see also Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

As was indicated earlier, our method is analogous to that used in the other papers (Bolger & Harvey, 1993; Lawrence & O'Connor, 1992) that have identified the use of the anchor and adjustment heuristic in judgmental time series forecasting using regression models. In addition to what was done in these papers, we also made many checks to rule out possible alternative anchors from among a very large number of possible candidates. Moreover, the use of regression models (or policy capturing) is a well-established method for identifying the mechanisms underlying judgment (e.g., see Carroll & Johnson, 1990). Indeed, the entire literature on the Brunswik lens is founded on this approach (e.g., see Cooksey, 2008).

One concern that may be associated with the use of multiple-choice tests in research such as this is that there might not be sufficient variation in the possible scores, or the nature of the distribution of scores may not reflect the patterns that are seen when measuring performances in other domains. However, in the cases discussed in this paper, we found the scores to be close to a normal distribution, which is a distribution that is commonly found to approximate performance scores in a wide range of contexts. Moreover, there are many practical situations where the possible variation in performance scores is much less than those found in this study (e.g., student feedback scores for lecturers are often measured on a scale of one to five, while the research performances of staff and departments in the UK are measured on a scale of one to four).

Also, our analysis was limited to forecasts of test scores, rather than of percentiles. This was because we wanted our participants to be forecasting a variable that had personal

[1] Recall that the estimated mean anchors in Table 2 were originally based on the assumption that the unanchored forecasts were unbiased. However, virtually the same mean anchor estimates apply when any bias is taken into account. This can be seen by substituting $A = (U - a)/b$ into the models in Table 2, using the estimated values of a and b displayed in Table 3. This arises because G is not correlated with A on any of the tests, meaning that its omission has little effect on the estimated regression coefficient for A if it is regarded as a missing variable in the models in Table 2. This can also be seen in the similarity of the values of β in Table 2 and β_2 in Table 3.

consequences. It is their mark that determines whether they pass or fail, as well as their degree classification. Percentiles are of little or no relevance to students, and they are never informed of their performances in terms of percentiles. It seems likely that percentiles would be more difficult to forecast, because not only must individuals forecast their own performances, they also have to determine how they will compare with those of others in their cohort.

One must note that the way in which questions are framed can influence individuals' responses, as has been demonstrated widely in the literature (see for example Yeung, 2014). However, we do not believe that our study is a victim of these effects, as the question the participants are asked is simply what mark they expect on a test. We have not complicated the task by asking overly elaborate questions, or asking a series of questions that might have clouded the respondent's judgement.

Of course, in some contexts, an individual's performance will also depend on external factors such as competition and teamwork. For instance, one's performance in a game of basketball will be influenced by both the obstacles put forward by the competition and the quality of other team members' performances. By isolating the participant in our experiment, we control for these factors in order to enable us to understand individuals' abilities to forecast their own performances when they are completely their own responsibility and depend solely on their own skills and knowledge.

Nevertheless, our research could be taken forward by looking at how individuals forecast their own performances under conditions of competition and teamwork. This would enrich our results by enhancing our understanding of the roles of competition and teamwork in altering one's beliefs (as compared to our findings in this paper) about one's own performance. In such cases, additional theories might need to be introduced to explain this part of the variation in forecasts. Given that performance is rarely judged in isolation but generally depends on other individuals, we believe that this type of research would be invaluable.

7. Conclusions

The use of the anchoring and adjustment heuristic provides a parsimonious explanation for the general tendency for individuals to produce regressive forecasts of their future performances. It does not require different theories for good and poor performers, nor does it leave unexplained the question of why we might expect a slope of less than one when we regress the forecasts on the actual scores. Of course, our analysis does not prove that other researchers' explanations are wrong. Indeed, the biases in these forecasts may be a result of several factors, including anchoring. Nevertheless, we believe that attempting to model the process by which people make their forecasts adds a new dimension to the debate.

Inevitably, our findings have a number of caveats. Our analysis is based on three cohorts of students who are taking tests in two quantitative subjects at a single university. Moreover, as was the case in most previous studies, these students were attending a leading university (Krajc & Ortmann, 2008) and taking a course that had very demanding entry requirements, meaning that few if any of the participants could be described as unskilled or ignorant.

Despite these caveats, there are a number of practical forecasting situations where an awareness of the anchoring effect and an ability to mitigate it may improve people's forecasts of their own performances. For example, this would be useful in group forecasting situations, where more accurate self-rated levels of expertise would provide an improved basis for weighting group members' forecasts.

Further research could usefully explore whether our findings also extend to people's predictions of the quality of their judgmental forecasts in areas such as sales or cost forecasting. For example, forecasters in an industry might perceive the typical MAPE to be 20% or a typical absolute error to be 300 units. These values may act as anchors, thus replicating the effect that we have found (although lower values here would signify better performances). This may lead to relatively poor forecasters underestimating their likely forecast errors, with the reverse being true for relatively good forecasters. As a result, insufficient safety stocks are held to cover for the expected forecast errors of the poorer forecasters, while excessive safety stocks are held to cover for the expected errors of relatively accurate forecasters. We believe that our study is potentially relevant here, as the anchoring and adjustment explanation has been found to apply across a wide range of different contexts.

References

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*, 394–400.

Armstrong, J. S. (2001). *Principles of forecasting*. Boston: Kluwer Academic Publishers.

Bolger, F., & Harvey, N. (1993). Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology*, *46*, 779–811.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes*, *88*(2), 719–736.

Burson, K. A., Larrick, R. P., & Kayman, J. (2006). Skilled or unskilled, but still unaware of it. How perception of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*, 60–77.

Carroll, J., & Johnson, E. J. (1990). *Decision research: a field guide*. Newbury Park: Sage.

Cervone, D., & Peake, P. K. (1986). Anchoring, efficacy, and action: The influence of judgment heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology*, *50*, 492–501.

Choplin, J. M., & Tawney, M. W. (2010). Mathematically modeling anchoring effects. In *Proceedings of the thirty-second annual conference of the cognitive science society* (pp. 501–506).

Clayson, D. E. (2005). Performance overconfidence: metacognitive effects or misplaced student expectations? *Journal of Marketing Education*, *27*, 122–129.

Cone, J., & Dunning, D. (2011). *Does genius go unrecognised?* Cornell University, Unpublished manuscript.

Cooksey, R. W. (2008). *Judgment analysis*. Bingley Emerald.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*, 118–121.

Domingos, P. (1999). The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery*, *3*(4), 409–425.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, *45*, 1007–1016.

Dunning, D. (2013). The problem of recognizing one's own incompetence: implications for self-assessment and development in the workplace. In S. Highhouse, R. S. Dalal, & E. Salas (Eds.), *Judgment and decision making at work* (pp. 37–56). Routledge.
Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organisation Behaviour and Human Decision Processes*, *105*, 98–121.
Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*, 465–485.
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.
Fagot, B. I., & O'Brien, M. (1994). Activity level in young children: Cross-age stability, situational influences, correlates with temperament, and the perception of problem behaviors. *Merrill Palmer Quarterly*, *40*, 378–398.
Felson, R. B. (1981). Ambiguity and bias in the self-concept. *Social Psychology Quarterly*, *44*, 64–69.
Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). London: McGraw-Hill.
Kennedy, E. J., Lawton, L., & Plumlee, E. L. (2002). Blissful ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education*, *24*, 243–252.
Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unware? An alternative explanation. *Journal of Economic Psychology*, *29*, 724–738.
Krueger, J., & Mueller, R. A. (2002). Unskilled unaware or both: The better than average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*, 180–188.
Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, *77*, 221–232.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own competence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.
Kruger, J., & Dunning, D. (2002). Unskilled and unaware — but why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology*, *82*, 189–192.
Kubovy, M., & Psotka, J. (1976). The predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experiment al Psychology: Human Perception and Performance*, *2*, 291–294.
Kunkel, E. (1971). On the relationship between estimate of ability and driver qualification. *Psychologie und Praxis*, *15*, 73–80.
Larwood, L., & Whittaker, W. (1977). Managerial myopia: Self-serving biases in organizational planning. *Journal of Applied Psychology*, *62*, 194–198.
Lawrence, M., & O'Connor, M. (1992). Exploring judgmental forecasting. *International Journal of Forecasting*, *8*, 15–26.
Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: its effectiveness and biases. *Journal of Behavioral Decision Making*, *8*, 149–168.
Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin and Review*, *1*, 126–129.
Miller, T. M., & Geraci, L. (2011). Unskilled and aware: reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*, 502–506.
Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). *Managing self-confidence: theory and experimental evidence*. Working Paper No. 17014. Cambridge, MA:NBER.
O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. *International Journal of Forecasting*, *9*, 163–172.
Pornprasertmanit, S., & Little, T. D. (2012). Determining directional dependency in causal associations. *International Journal of Behavioral Development*, *36*, 313–322.
Poundstone, W. (2014). *How to predict the unpredictable*. London: Oneworld.
Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attributional processes. *Journal of Experimental Social Psychology*, *13*, 279–301.
Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, *15*, 353–375.
Russo, J. E. (1978). Eye fixations can save the world: A critical evaluation and a comparison between eye fixations and other information processing methodologies. *Advances in Consumer Research*, *5*, 561–570.
Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, *39*, 85–100.
Sheldon, O., Ames, D., & Dunning, D. (2011). *Self-assessments of emotional intelligence*. Rutgers University, Unpublished manuscript.
Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and improvement of forecasts. *Journal of Forecasting*, *13*, 579–599.
Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*, 437–446.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
Yeung, S. (2014). Framing effect in evaluation of others' predictions. *Judgment and Decision Making*, *9*, 445–464.

Sheik Meeran is an Associate Professor (Senior Lecturer) in decision analysis in the School of Management, University of Bath, UK. He is a Fellow of the Higher Education Academy of UK, and did his undergraduate degree in engineering in Madurai University, India, passing the same with a special Honours and University First rank. After graduation, he worked in industry for nearly 10 years, spending a substantial portion of this time in production management. He then continued with his Masters and Doctoral degrees in Cranfield, UK, in the domain of manufacturing management. Since then, he has been teaching at various UK universities. In addition to successfully supervising more than ten Ph.D. candidates, he has published in many reputable international journals, especially in the fields of CADCAM integration (feature recognition), job scheduling and forecasting. One of his papers has been cited nearly 500 times since its publication in 1999. Currently, one of his main research interests is in the field of new product forecasting and self-performance forecasting.

Paul Goodwin is Emeritus Professor of management science in the School of Management, University of Bath, UK. He has degrees from Liverpool and Warwick universities and a Ph.D. from Lancaster University. His research interests focus on the integration of management judgment and statistical methods in forecasting and decision making. He is the co-author of *Decision analysis for management judgment* (4th edition) (Wiley) and co-editor of *Forecasting with judgment* (Wiley). He has published over 50 papers in international journals, and is an Editor of the *International Journal of Forecasting* and a member of the editorial boards of the *Journal of Behavioural Decision Making* and *Foresight: the International Journal of Applied Forecasting*. In addition, he is a former Director of the International Institute of Forecasters.

Baris Yalabik is Lecturer (Assistant Professor) in operations and supply management in the School of Management, University of Bath, UK. His research examines sustainable approaches to the design and management of production systems. He also has research interests around teaching and learning in management.