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 choice, 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 performance on tasks can also be important to those involved in judgmental forecasting itself. For example, a tendency to over forecast one’s future performance in, sales forecasting, might lead to unresponsiveness to advice and feedback (e.g., Bonaccio and Dalal, 2006; Dunning, 2013; Lim and 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 a person to overweight their 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 people to discount potentially valuable inputs that they could make to the forecasting process. (though see Rowe and Wright (1999) found the evidence linking confidence in one’s forecast and willingness to change to be inconsistent). These potential problems would particularly apply to group-based forecasting methods that require group members to explicitly self-rate their expertise (e.g. DeGroot, 1974).

A number of studies have investigated how accurate people are at forecasting their performance on tasks, tests or examinations (Clayson, 2005, Miller & Geraci, 2011, Kennedy et al., 2002, Burson et al., 2006, Kruger & Dunning, 1999 and Krueger & Mueller, 2002).

The results have varied between those that found no correlation between predicted performance and actual performance to those that found a significant correlation. However, even where there is a significant positive correlation, a common finding is that, on average, relatively poor performers tend to over forecast their performance while high performers tend to 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 and Mueller, 2002). In this paper we adopt a judgmental forecasting perspective to suggest an alternative explanation for this tendency and we 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 first review the literature relating to this topic before developing a theoretical model to represent the forecasting process. We then present the analysis of data from six in-course multiple-choice tests of statistical and forecasting knowledge. This enabled us to model the process used by people to forecast their own performance under conditions where 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 performance in these tests. A key advantage of using multiple-choice tests in this research is that the scores achieved are determined objectively. The use of marks on 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 to which subjectivity contributes to the student's mark is far less than in many other forms of performance assessment.

Literature review

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

There may be a number of factors that lead to individuals over forecasting their test performance. One well known phenomenon is the “above average effect” where most people perceive that their skills are above average. This has been observed in areas ranging from football (Felson, 1981) to business management and leadership (Larwood and 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 directly explain why the forecasts are regressive in that the optimism is only associated with low performance. Nor does it directly explain 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 this. A more elaborate explanation is provided by Kruger and Dunning (2002). They 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 that they are incompetent (also see Ehrlinger et al., 2008). More recently, Dunning (2013) has presented several examples of unskilled individuals being unaware of their lack of skill, and mentions implications for organisations such as the difficulties of recognising expertise in groups (Cone and Dunning 2011; Bonner et al 2002), reluctance to seek advice (Bonnacio and Dalal 2006), and evaluating feedback for the purpose of self-improvement (Sheldon et al 2011; Mobius et al 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 et al., 1977) and argue that the high achievers assume that their peers are as skilled as they are. Hence they tend to assess their ability as being closer to the middle of the range of performance 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 reasonable accurately, other studies have found this bias also occurs with forecasts of scores (Clayson, 2005 and Miller & Geraci, 2011).

Later studies have reconsidered the notion that poor performers are unaware of their limitations. Of course, in this context the term 'limitations' can refer to a number of separate abilities. These include: i) lack of ability relating to the skill that is being assessed in the test; ii) lack of ability to make a self-assessment of one's skill in this domain iii) lack of ability to convert (ii) into a forecast of one's score and iv) lack of ability to appraise the likely accuracy of this forecast. Miller and Geraci (2011) addressed the fourth of these and found that poor performers exhibit more uncertainty about their forecasts. They found that, while poor performers tended to over forecast their grade, they also attached lower confidence ratings to their forecasts suggesting that they are very much aware of their inability to make accurate assessments of their skill. A reason for these inaccurate assessments is suggested by Krajc and Ortmann (2008). They point out that most studies have been conducted in elite educational institutions where there is a bunching of grades. They 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, Schlosser et al (2013) find no evidence that this explanation, is behind 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 regression analysis. When the forecasts and actual scores on a test are imperfectly correlated 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 1. Given that perfectly accurate forecasts would be represented by a line where a = 0 and b=1, the two lines will cross so that the poor performers tend to over forecast on average and the high performers under forecast . (Alternatively, if the lines do not cross within the bounds of scores determined by the test, the tendency to over forecast will decline the higher an individual's actual performance). Because of the above average effect the intercept of the line 'a' will tend to be relatively high (see figure 1) so that the average level of under forecasting will be less than the average level of over forecasting.

(**please insert figure 1 about here**)

Note that regression factors will only account for the phenomenon when conditions lead to the line's slope being less than 1. Where the forecasts vary more than the actual scores (a situation that is entirely possible in judgmental forecasting (O'Connor et al., 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 much be smaller than those of the forecasts and, depending on the correlation between the forecasts and scores, the slope of the line could exceed 1. Thus, if regression is used to explain the phenomenon, it simply raises a new question as why the slope of the line should be expected to be less than one.

Burson et al. (2006) also argue that metacognitive factors do not lead to the bias. They found when people were asked to predict the percentile for their performance on a series of tests the best and worst performers performed very similarly in their predictions of performance when the task was of moderate difficulty. Moreover, on very difficult tasks the best performers actually produced less accurate forecasts than the worst performers.   Burson et al. (2006) proposed that inaccuracy in the percentile forecasts was due to a combination of noise (for example 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 feedback on performance.

Making a prediction of one's mark on a test is essentially a judgmental forecasting task. Judgment is used to integrate the perceived effect of available cues (these may include previous test marks, a perceived 'norm' mark on the test or one's perceived level of effort in preparing for the test) to forecast the value of an uncertain quantity, given that this value will not be known until the future (Armstrong, 2001, p790). In some aspects the task parallels that of a sales person forecasting the sales 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 their performance and poor performers to over forecast . This literature can potentially yield a number of new insights into the causes of the bias and how 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 assessing the components of skill of the judgmental forecaster (Stewart and Lusk, 1994). High correlations between forecasts and outcomes are often quoted, but they can be misinterpreted (Kruger and 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 available information (e.g. Lawrence and 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 and 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 take into account the mean of the series (Bolger & Harvey, 1993 and Lawrence & O'Connor, 1992). A consequence of using the anchoring-and-adjustment heuristic is that the adjustment from the anchor is usually insufficient to reflect the implications of the new information. The anchor can be particularly powerful when it is self-generated (Cervone & Peake, 1986 and Strack & Mussweiler, 1997).

At least one study has hinted that people may use anchoring and adjustment when forecasting their test mark. 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 already 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. They, 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 make perfect assessments of their own ability, the regressive forecasts that have been encountered in earlier studies would still be observed.

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 an estimate, U, that the person would have made had they not seen the anchor (e.g. see Choplin and Tawney, 2010). That is:

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

where w = a weight and k is noise.

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

$$U = a + bA + i \qquad \text{where A is the true mark and i is noise} \qquad (2)$$

and a and b are bias parameters reflecting the fact that students may have wrongly estimated factors such as the difficulty of the test. If b= 0 it 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:

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

where j is noise representing different susceptibility to the anchor.

Thus:

$$F = (1-w)G + w\,[a + bA\,] + e \qquad \text{where } e = w\,i + j$$

$$= (1-w)G + wa + wbA + e \qquad (3a)$$

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

$$F = \alpha + \beta A + e \qquad (4)$$

So from (3a) and (4):

$\alpha = (1-w)G + wa$ so $G = (\alpha - wa)/(1-w)$

also $\beta = wb$ so $w = \beta/b$

hence $G = \frac{1}{1-w}\left[\propto - \frac{\beta}{b} a\right]$ (5)

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

$$G = \frac{\propto}{1-\beta} \qquad (6)$$

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 (and we test for this later) equations (5) or (6) yield an estimate of the mean value of the anchors that they used.

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 the 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 during the autumn (or fall) semester.

Please insert table 1 about here

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 biases associated with judgmental forecasting were explored on the Forecasting course the tests took place before this topic was presented. No credit was given for participation. The students took multiple choice tests, ranging in time from 20 to 45 minutes, that were designed to assess their understanding of the concepts that had been covered on the course prior to the test. Before the test they were asked to forecast the mark they thought they would achieve. In tests 2 to 6 they were then asked to forecast the mean mark that they thought the cohort would achieve. They were assured that their forecasts would have no influence on the mark that they would be awarded and that their forecast would be treated as confidential.

Analysis

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

Please insert table 2 about here

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 below a mark of $\alpha/(1-\beta)$ 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 forecasts of the cohort mean mark produced by the students. The largest differences appear for tests 2 and 6 where the students had no experience of taking these tests, but even these differences are small. Also, in all cases the average forecasts of the cohort mean mark are found within the range 70.8% to 74.14% of the maximum mark. 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 their individual marks forecast. We have not yet established that the unanchored forecasts were unbiased but this demonstrates the possibility that students can have an unbiased expectation of their performance 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 above results provide a prima facie case that the students were on average, 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 to both their prediction of the cohort mean (G) and their actual mark (A) yielding models of the form:

$$F = \beta_0 + \beta_1 G + \beta_2 A + e \qquad (7)$$

Comparing this with (3a):

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

It can be seen that:

β_0 =wa so a = β_0/w

β_1 = (1-w) so w = 1- β_1

β_2= wb so b = β_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 the coefficient for the individual's prediction of the cohort mean mark is significant at least at the 5% level. 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.

Please insert table 3 about here

Note that tests, 2 and 6 were 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 the 'false consensus' explanations or simply the inability of the students 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 the regressive effect. In particular, the results for test 2 suggest that a major factor was the students' tendency to underestimate the test's difficulty. 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 so, on average, almost all the students would be expected to over forecast their mark. Test 4 had higher mean marks than the other tests and so it may have been easier than the students expected. However, even in the case of these 3 tests the results suggest a further biasing effect as a result of anchoring.

In summary, our models provided evidence of anchoring in *every* case 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 make 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 they are forecasting tomorrow's midday temperature in Seattle, the anchor is clearly biased but the individual can still be using the anchor and adjustment heuristic when they make their forecast. Thus the existence of a biased anchor does not mean that anchoring and adjustment was not being employed. Even where other theories suggested in the literature apply, such as the metacognitive explanation, anchoring may lead to additional biasing effects.

Discussion

The above analysis 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 and 3) why would values distributed around 72% act as anchors? In this section we will address 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 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 indicated in table 2, this coincided with values close to 72% of the maximum marks, suggesting that other points did not act as a common anchor.

Fitting models of the form shown in (7) with G representing (ii) to (v) above, in turn, always led to R^2 values that were much lower than those values (shown in table 3) where G equalled the predicted cohort mean (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 but instead were themselves anchored on the individual forecasts of the test mark. After all in Tests 2 to 6 the students were asked for their forecast of the cohort mean directly after they had made a forecast of their own individual mark. Tests for the direction of causality in the regression models based on coefficients of kurtosis were inconclusive (Pornprasertmanit and Little, 2012). However, there is some evidence that the cohort mean as the more likely anchor. First, in Test 1 the students were not asked for a forecast of the cohort mean before they took the test but the estimated mean of the anchors they used, shown in table 2, is very similar to that 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 marks forecasts and suggest that these were anchored on the predicted cohort means *before* these cohort mean predictions were formally elicited. [1] In addition, Kruger (1999) predicts that the 'above-average' effect will prevail in situations where 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 performance form the anchor. Yet table 1 shows that, in four of the five tests for which information is

[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 are displayed in table 3. This arises because G is not correlated with A on any of the tests so if it is regarded as a missing variable in the models in table 2, its omission has little effect on the estimated regression coefficient for A. This can also be seen in the closeness of the values of β in table 2 and β_2 in table 3.

available, the mean forecast of their mark made by individuals in our tests was below their forecast of the cohort mean so 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 space of 8 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. Poundstone, 2014, Duan, Gu & Whinston, 2008) so there may be a natural tendency to use values close to 70% as the starting value in estimation and forecasting. Also people most often choose 7 when asked to select a number from a 0 to 9 scale (Kubovy and Psotka, 1976) Indeed, the predictions of the cohort mean may, themselves, have been anchored 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 have been seen as feasible. It is also worth noting that there is a commonality between these cohorts is that their entry requirement to their degrees are the same and require 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 meant that we could only infer whether anchoring and adjustment was typically being used by modelling the average responses of a large sample of participants. Verbal protocol analysis would have offered an alternative approach (Epley and Gilovich, 2001). However, we note that application of it in our work has 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 could come into play so that the reported process might not reflect that which would otherwise have been naturally used. Moreover the act of providing a verbal account of one’s judgment process will itself use cognitive resources, thereby reducing those available to the judgment task and so 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 allows only small samples to be used because of its high resource demands. Indeed Carroll and Johnson (1990) argue “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” (also see Einhorn, Kleinmuntz and Klienmuntz, 1979).

As indicated earlier, our method is analogous to that used in the other papers (Bolger and Harvey 1993; Lawrence & O’Connor, 1992) which 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 made many checks to rule out possible alternative anchors from a very large number of possible candidates. Moreover, the use of regression models (or policy capturing) is a widely established method for identifying mechanisms underlying judgment (e.g. see Carroll and 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 be insufficient variation in the possible scores, or the nature of the distribution of scores may not reflect patterns that are seen when measuring performance in other domains. However, in the cases discussed in this paper we found that the scores were close to a normal distribution which is a distribution commonly found to approximate performance scores in a wide range of contexts. Moreover, there are many practical situations where variations in performance scores are much less than those found in this study (e.g. student feedback scores for lecturers are often measured on 1 to 5 scale, while in the UK the research performance of staff and departments are measured on a 1 to 4 scale).

Also, our analysis was limited to test scores, rather than forecasts of percentiles. This was because we wanted our participants to forecast a variable that had personal consequences. It is their mark that determines whether they pass or fail, as well as their degree classification. Percentiles have little or no relevance to students, and they are never informed of their performance in terms of a percentile. It seems likely that percentiles may be more difficult to forecast because the individual not only has to forecast their own performance, they also have to determine how their performance will compare with others in their cohort.

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

Of course, in some contexts an individual's performance will also be dependent on external factors such as competition and teamwork. For instance, one's performance in a game of basketball will be influenced by 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 to enable us to understand individuals' ability to forecast their own performance when this is completely their own responsibility and is solely dependent on their own skills and knowledge.

Nevertheless, our research may be taken forward by looking at how individuals forecast their own performances under competitive and teamwork conditions. This would enrich our results by understanding the role of competition and teamwork in altering one's own beliefs (as compared to our findings in this paper) about their own performance. Thus, perhaps additional theories would need to be put into play to explain this part of the variation in forecasts. Given that performance is not often judged in isolation but is dependent on other individuals, we believe this type of research would be invaluable.

Conclusions

The use of the anchoring and adjustment heuristic provides a parsimonious explanation for the tendency of people to produce regressive forecasts of their future performance. It does not require different theories for those who perform well and those who perform poorly, nor does it leave unexplained why we might expect a slope of less than 1 when we regress the forecasts onto the actual scores. Of course, our analysis does not prove that explanations suggested by other researchers 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 through 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 taking tests in two quantitative subjects at a single university. Moreover, as in most previous studies these students were attending a leading university (Krajc and Ortmann, 2008) and taking a course that had very demanding entry requirements so 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 being aware of the anchoring effect and being able to mitigate it may improve people's forecasts of their own performance. This would be useful, for example, in group forecasting situations where more accurate assessments of self-rated levels of expertise would provide an improved basis for weighting the forecasts of group members.

Further research could usefully explore whether our findings translate into 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 that the typical MAPE is 20% or a typical absolute error is 300 units. These values may act as an anchor so that the effect we found is replicated (although here lower values would signify better performance). This may lead to relatively poor forecasters underestimating their likely forecast error, with the reverse being true for relatively good forecasters. As a result insufficient safety stocks are held to cover for the expected forecast error of the poorer forecasters while excessive safety stocks are held to cover for the expected error 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.

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. and 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.*

Clayson, D. E. (2005). Performance Overconfidence: Metacognitive Effects or Misplaced Student Expectations? *Journal of Marketing Education, 27*, 122-129.

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

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

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.*Judgment and Decision Making at Work*, 37.

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.

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.

Ferraro, P. J. (2010). Know thyself: Competence and self-awareness. *Atlantic Economic Journal*, *38*(2), 183-196.

Gujarati, D.N. (1995*). Basic Econometrics* 3rd edition. 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.

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

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. & 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.

Lawrence, M. & O'Connor, M. (1992) Exploring judgmental forecasting. *International Journal of Forecasting, 8,* 15-26.

Larwood, L. & Whittaker, W. (1977). Managerial myopia: Self-serving biases in organizational planning. *Journal of Applied Psychology, 62*, 194-198.

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

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 & 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(0): 85-100.

Sheldon, O., Ames, D., Dunning, D. (2011). Self-assessments of emotional intelligence. Unpublished manuscript, Rutgers University.

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.