



BANK OF ENGLAND

Speech

Forecast errors

Speech given by

Ben Broadbent, External Member of the Monetary Policy Committee, Bank of England

At The Mile End Group of Queen Mary, University of London

Wednesday 1 May 2013

I would like to thank Alina Barnett, Adrian Chiu and Amardeep Parmar for research assistance, and I am also grateful for helpful comments from other colleagues, especially Andrew Blake, Nicholas Fawcett, Konstantinos Theodoridis and Martin Weale. The views expressed are my own and do not necessarily reflect those of the Bank of England or other members of the Monetary Policy Committee.

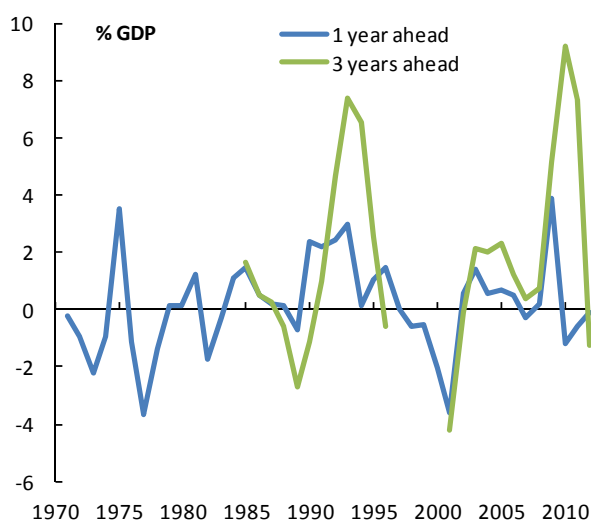
Forecast errors

There's a well-known joke about economic forecasting. Albert Einstein reaches the pearly gates of heaven and meets three people. He asks them all for their IQ. "190", says the first. "Oh good," says the great man, "we can talk about general relativity". When the second says "140", Einstein tells him he's looking forward to debates about the pros and cons of the nuclear deterrent. The third, peering at his feet and mumbling slightly, says "mine's only 50". "What will the budget deficit be next year?"

I feel I'm allowed to tell this joke because my first job, at the Treasury, was in a team responsible for forecasting the budget deficit. And if the joke is meant as comment on the accuracy of such forecasts (or rather the lack of it), my early experience does not, regrettably, provide much of a defence. The first Budget forecast I went through was in 1989. That year's "Red Book" predicted that, three years later (in 1992-93), the government would be running a small financial surplus. What we got, as things turned out, was the (then) largest peacetime deficit on record.

One wonders what Einstein – whose theories generate predictions that, even at relative speeds of 100,000 miles per hour, differ from Newton's by one part in 10 million – would have made of this. Certainly economic commentators of the time were unimpressed. And it's not been of much comfort that we've since done worse. If the failure to foresee a relatively contained and short-lived recession in the early 1990s did little to enhance the reputation of economic forecasting at that time, the failure to anticipate the global financial crisis of 2008 – which, as you can see from Chart 1, led to an even bigger error in the forecast for UK government borrowing – has diminished it still further. Famously, even the Queen saw fit to put the profession on the spot, asking economists publicly why no-one had seen the crisis coming.

Chart 1: Public borrowing, outturn less forecast



Source: HMT

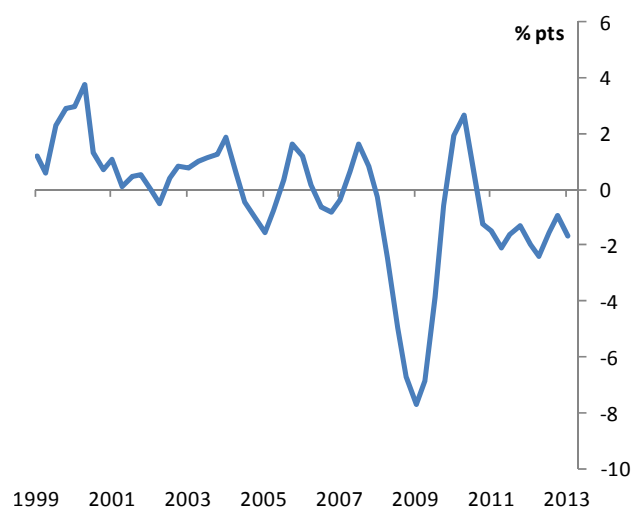
Laws and forecasts are different things and it's unfair to compare them directly. It's true that economics doesn't have empirical theories as precise as those in the natural sciences. Nor can economists test their ideas as accurately. But even physical laws cannot provide us with wholly accurate forecasts, whether about the position of individual subatomic particles a moment from now or next week's weather ("prediction is difficult", said Niels Bohr, one of the founding fathers of quantum theory, "especially about the future").

However much we dislike them, therefore, forecast errors are inevitable. Indeed, without them we wouldn't even need forecasts. A world without forecasting errors wouldn't have the need for any financial assets either – or none, at least, whose prices ever changed¹. Nor would we ever experience financial crises. In fact, one answer to the royal question – “why didn't we see the crisis coming?” – is that, had it been easy to do so, we would not have had one, as people would have taken steps to avoid it. We only get to observe the crises that people didn't foresee.

This is hardly a complete, or very satisfying, answer. The events of 2008 may not have been easily predictable. But if it's true that being aware of the risk of a financial crisis means people take steps to limit its costs, then surely our duty to look for advance warning signals is all the greater. And the fact that forecasting errors are inevitable does not mean – clearly – that all forecasts are equally good or that none is beyond improvement.

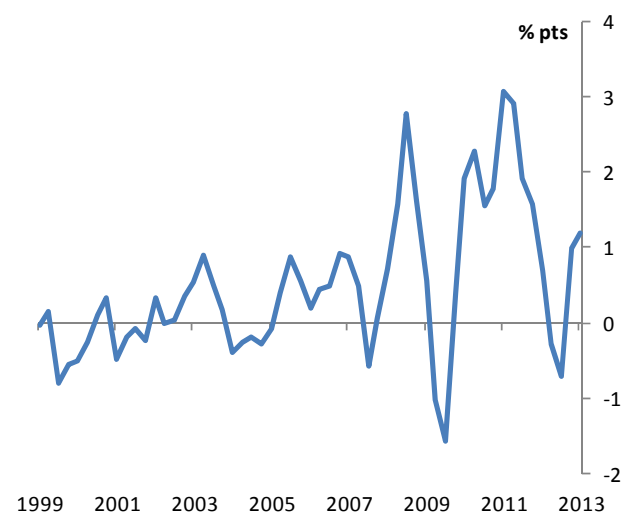
In his review of the MPC's forecasting capability, published last autumn, David Stockton, formerly a director of research at the US Fed, said “the MPC's recent forecasting performance has been noticeably worse than prior to the crisis and marginally worse than that of outside forecasters. [Its] forecast errors have been characterised by persistent over-prediction of [economic] growth and under-prediction of inflation”. These errors (this time relative to forecasts just one year earlier) are plotted in Charts 2 and 3.

Chart 2: GDP growth, out-turn less MPC's mean forecast



Source: Bank of England and ONS

Chart 3: Consumer price inflation, out-turn less MPC's mean forecast



Note: relates to RPIX inflation between 1999 and 2004, CPI inflation thereafter.

Source: Bank of England and ONS

¹ Financial assets serve two purposes – they facilitate risk sharing and allow for inter-temporal trades (between young and old, for example, or patient and impatient people). So they would still have a purpose in a riskless world, but any change in price would be deterministic and predictable.

In addition, while he found the Bank's explanation of these findings to be persuasive "in broad terms", Stockton also said "[its] narrative may not fully explain the persistence of these recent errors...[which] could reflect some inertia imparted by the forecast process or...problems with the paradigm underlying the Bank's forecasts." There has also been criticism of the forecasts – most of it a good deal more trenchant than this – from domestic commentators.

The Bank of England has already given a detailed – and very positive – response to the Stockton Review and its recommendations² and I do not intend to add to that here. Nor, except in the broadest possible terms, will I attempt to analyse the MPC's forecasts or offer a blow-by-blow explanation for the recent record. What I do want to do, however, in the light of the recent criticism, is to reiterate that the mere existence of "errors" – even the word is something of a misnomer – is not, in and of itself, evidence that something is wrong, or even improvable.

Suppose you divide these errors, by their source, into three: known unknowns, forecasting mistakes and unknown unknowns (you might call this the "Rumsfeld Classification"). In the first bucket belong things like changes in oil supply, fluctuations in harvests or political surprises. These things cannot be predicted in advance. But they are all events with which we're familiar – they crop up regularly in our datasets – and we can therefore make reasonable attempts both to model their effects and to allow for the risk that they occur in future. These "known unknowns" are what give rise to the MPC's "fancharts"³.

The second category might include all sorts of things: a failure to choose the best possible model of the economy, excluding some relevant variable or any number of other hazards⁴. The important point is that, to qualify as a genuine fault, it should be reasonably identifiable at the time the forecast is being made. Failings that appear only with the benefit of hindsight belong in one of the other two categories⁵. Note too that, if correctable mistakes are possible, and regularly made, then one should probably expect some forecasts to do measurably better than others.

The third category is the most problematic. It has been populated in the past by black swans, round (as opposed to flat) earths and a whole host of things that weren't even countenanced beforehand. More prosaically, it also includes what economic modellers call "structural breaks": shifts in things that, in our models, we were treating as unchangeable constants.

² "Response of the Bank of England to the Three Court-Commissioned Reviews", at http://www.bankofengland.co.uk/publications/Documents/news/2013/nr051_courtreviews.pdf

³ The fancharts, and this first category, will also include the effects of sampling error around the estimated parameters in the forecasting model. See Elder et al. (2005).

⁴ Perhaps the MPC's failure to allow sufficiently for the "pass-through" from the depreciation of sterling in 2008, when forecasting CPI inflation over the following couple of years, falls into this category (Dale (2011)). If so, it's a mistake I and others made too.

⁵ To be precise, it would not be right to classify something as a "mistake" if, as an all-knowing and fully informed critic, you found no problem with it at the time. It is quite possible for information about genuine and correctable mistakes to emerge only after the event. But that is still something you would infer from an examination of the methodology itself, not the fact that the resulting forecast was "wrong".

It is impossible to make this classification entirely cleanly: one person's economic "shock" may be another person's forecast failure. So a forecaster who commits genuine mistakes may well be tempted to blame the resulting forecast error on something else, particularly if it involves a technical matter that an uninformed outsider cannot observe or understand directly. Knowing this, the outsider would naturally tend to view with scepticism protestations that errors are unavoidable.

But I nonetheless believe that there is probably too much scepticism: people are too inclined to put into the second bucket what belongs in one of the others. After a short description of how noisy many economic series (in this case GDP growth) really are, I'll make two broad points in this regard.

First, at least under some circumstances, distinguishing between economic models, and their forecasting performance, can take a long time. The greater the degree of true randomness in the world, the rarer the event being modelled and the stronger one's prior belief about the structure of the economy, the longer it takes to be confident that any given forecast process is flawed.

Yet – and this is the second point – we are, all of us, genetically under-endowed with the patience it can require to make these distinctions. We are instead, as the psychologist Daniel Kahneman puts it, "machines for rushing to judgement", biased judgement at that. We are naturally too inclined to see structure in what is actually random. We are also too inclined to view a forecast "error" as precisely that: someone's mistake.

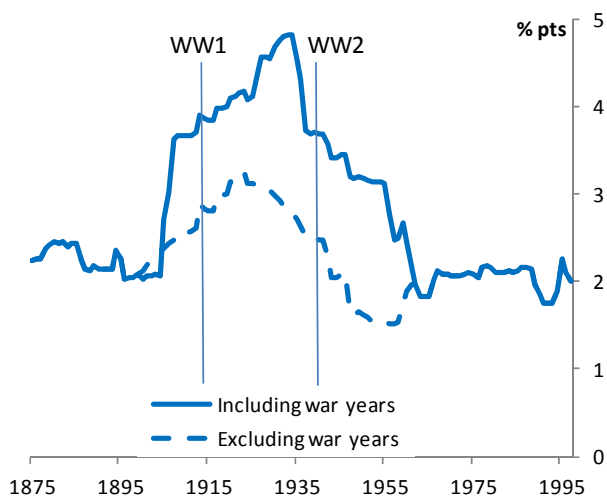
GDP growth: more noise than predictable signal

In a recent report criticising the MPC's forecasting performance (and before getting the gloves off), one domestic commentator made the concession that "we cannot expect forecasts to be 100% accurate". Quite so. In fact, in most cases we'd be happy if we got close to half way there.

Over the fifteen-year period for which we have a consistent set of macroeconomic forecasts, the standard deviation of annual GDP growth has been 2% points or so (growth has been within that margin of the sample average around two-thirds of the time). All twenty year-ahead growth forecasts contain some information, in that they are correlated with realised growth. But they do not contain much: of that 2%-point standard deviation, only 30%-40% of it (0.6% - 0.8% points) has on average been anticipated by any of the forecasts⁶. Thus even when we look only a year ahead, the unpredicted component in annual GDP growth – the "noise" – has been significantly greater than "signal" we're able to extract from the various economic indicators, and on average close to twice as big.

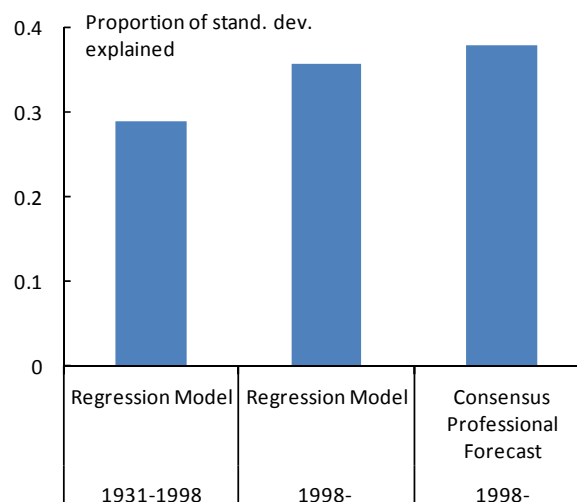
⁶ The share of the variance of output growth explained by the forecast is its $R^2 = 1 - \text{MSE}/\text{Var}(y)$, where MSE is the mean squared error and $\text{Var}(y)$ the sample variance of growth. We define the share of the standard deviation as $\sqrt{R^2}/[\sqrt{R^2} + \sqrt{(1-R^2)}]$ and the signal:noise ratio is $\sqrt{R^2/(1-R^2)}$.

Chart 4: Variability of annual GDP growth



Source: Bank of England

Chart 5: Variation explained by simple model and professional forecasters



Source: HMT and Bank of England calculations

These forecasts only go only to the late 1990s, since when there have been significant shifts in the world economy and a huge financial crisis. So perhaps we're judging them over a period in which the economy has been unusually volatile and accurate prediction unusually difficult. But the economy has always been volatile (Chart 4 plots the standard deviation of GDP growth over 30-year rolling periods and, even if you strip out the war years, it has never fallen much below that 2%-point figure). And although we do not have formal published forecasts for these earlier periods, a rolling sequence of simple regression models, driven by things you might think relevant for short-term growth⁷, actually seems to perform marginally less well in the decades prior to 1998 (we have data from 1930) than it would have done since (Chart 5).

On the face of it, it seems hard to generate year-ahead forecasts of UK economic growth that explain as much as one half the variation in UK economic growth, let alone the "100% accurate" benchmark cited by the critic.

More noise means weaker statistical tests

Some of you may conclude from this that all forecasters are poor and all should do better. And one hopes, of course, that as we learn more, and as estimation techniques and economic theory get refined, we can indeed reduce the size of these errors, particularly the most extreme misses. But we also have to accept that, to a significant extent, many objects of interest, including GDP growth, are genuinely unpredictable, comprising at least as much noise as signal. And that, in turn, means that, unless you have many years of

⁷ The model regresses growth on a number of things you might think relevant for short-term growth: the recent history of growth itself, dummy variables for war period and changes in short and long-term interest rates, the exchange rate, equity prices, oil prices and government spending. Everything is lagged at least once, to ensure our regression doesn't include any information that a forecaster wouldn't have access to at the time. For the same reason, these are "out of sample" forecasts: the prediction for growth in 1980, for example, is generated by a regression run on data from 1950 to 1979, that for 1981 by a 1951-80 regression and so on.

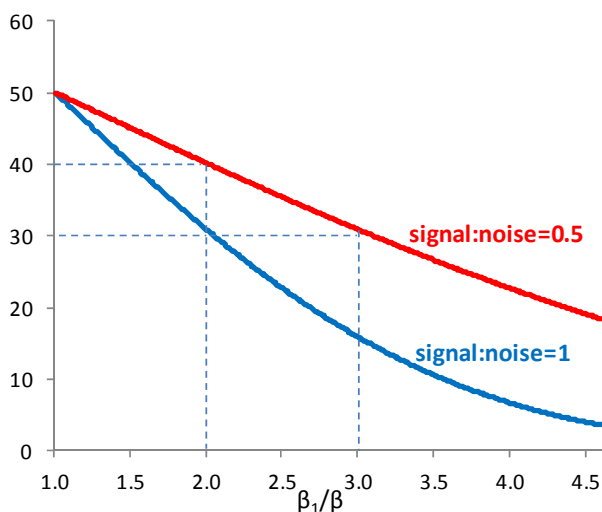
data to work with, you have to be careful about assessing and comparing forecasts. The greater the noise in a series, the more often a bad forecast will outperform a good one, and the harder it is to tell them apart. Let me illustrate this using a very simple simulation. Suppose we're interested in some variable Y , and that Y is generated by another variable X – which, though unpredictable at longer horizons, we learn of one period in advance – and a random noise term ε , which we do not. The precise impact of X on Y depends on some fixed number β :

$$Y = \beta X + \varepsilon$$

However, while we can all see X , we're collectively less sure about β , the degree to which it actually influences Y . Suppose, for the sake of argument, that you (the audience) and I both have to forecast Y , but while I know the true β you lot have got it all wrong and are using some $\beta_1 \neq \beta$ (I'm allowed to have it this way round as I'm the one giving the speech.)

The first thing to say is that, with any significant noise in the system, you will sometimes make smaller prediction errors than me, even though you've got the wrong model. In fact, if there's enough noise in the system, your model can be wrong by a wide margin and still, quite frequently, outperform my "perfect" (ie un-improvable) forecast.

Chart 6: Frequency with which bad model outperforms good



Each line in Chart 6 plots the likelihood, for a given signal:noise ratio (marked alongside the line), that the wrong model does better. The horizontal axis measures proportionately how far apart the two are: a reading of "2" indicates that β_1 is twice β , "3" that β_1 is three times β and so on. So if the bad model uses $\beta_1 = 2\beta$ – one might say you're "100% wrong" – and if the signal:noise ratio is 1 then, reading along the blue line, we can see that you can nonetheless expect to do better than me roughly one time out of three⁸.

⁸ I've assumed mean zero, normal distributions for both X and ε . Because they're symmetrically distributed, the probability that the bad model outperforms the good depends only on $|(\beta_1 - \beta)/\beta|$: it's the same whether β_1 is larger or smaller than β (by a given proportionate amount). So the continuation of Chart 7 to the left of the origin is just a mirror image of what's drawn here. The fact that X has a zero mean is important: if it did not, then the wrong model would have a persistent bias and the difference between β_1 and β would show up more quickly. But I take it as read that, in the real world, all forecasting models are made to fit sample means. Any differences would therefore be apparent only in higher moments of the data.

If the signal:noise ratio is only one-half (closer to the case of the average post-1998 GDP forecast) then even a “200% wrong” model ($\beta_1 = 3\beta$) will attain that one-in-three success rating. And the less (but still) wrong $\beta_1 = 2\beta$ model will on average beat the best forecast more than 40% of the time.

That’s still less than one half and the bad model will eventually get found out. But the key word is “eventually”: the closer the performance of the two models – as it will be the noisier the underlying data – the larger the sample you need to distinguish them. Otherwise, the gap in performance, such as it is, gets dominated by sampling error: what looks like superiority is more likely just good luck.

You can see the importance of sample size in Charts 7 and 8. Chart 7 picks one particular bad model, $\beta_1 = 2\beta$, and plots the likelihood that you can reject it (i.e. that its underperformance relative to the true model becomes statistically significant) against the size of the sample. The blue line, calculated for a signal:noise ratio of 1, tells you that you don’t need many observations – 9 or 10 – to have a better-than-evens chance of making a reasonably clear distinction between the two. At that point, the sampling error is already small enough that the outperformance of the good model will probably be statistically significant.

But the noisier and less predictable the underlying series the more data you need. When the signal:noise ratio falls to a half (the red line) you need a sample size of 30 to be reasonably confident of distinguishing good from bad. And if the bad model is closer to the truth ($\beta_1 < 2\beta$) the critical sample size will be that much higher, unsurprisingly. Again for two different signal:noise ratios, Chart 8 plots the critical sample size against the distance between the two models. For the noisier of the two series (the red line), our simulated model can be “75% wrong” ($\beta_1 = 1.75\beta$) and still, for sample sizes up to 100, have a better-than-evens chance of surviving a forecast comparison with the true model.

Chart 7: Probability that false ($\beta_1 = 2\beta$) model is rejected (at 5% significance)¹

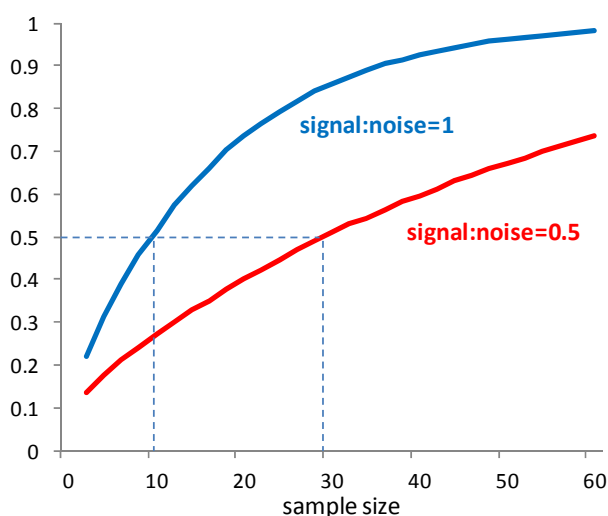
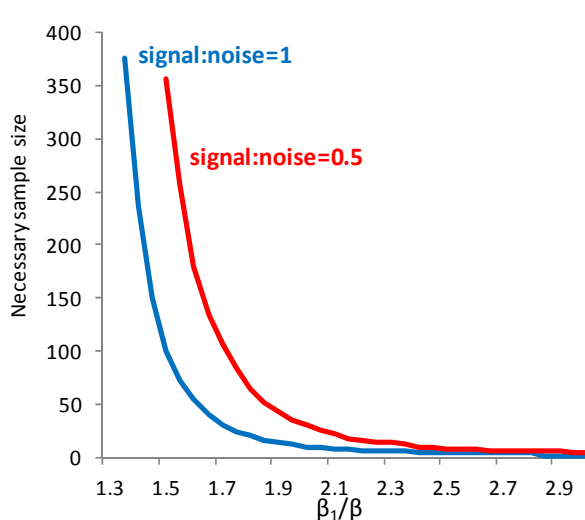


Chart 8: Sample necessary to make rejection more likely than not



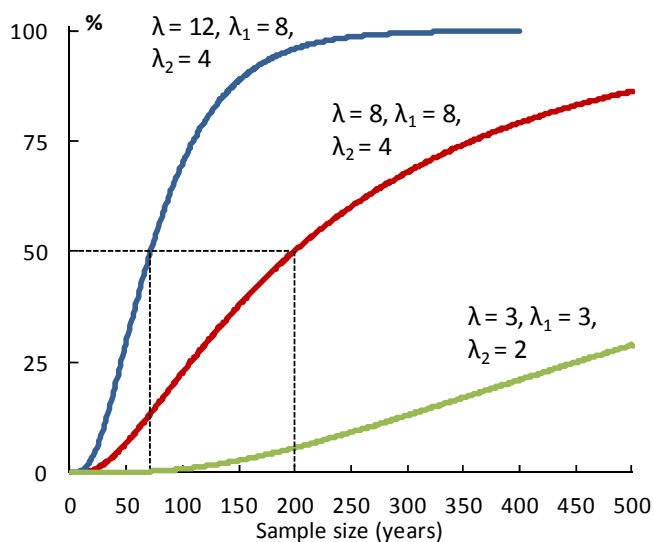
As we've seen, across the 20 separate organisations that make year-ahead forecasts of GDP, we currently have only fifteen years of data. Of the 190 possible pair-wise comparisons these 20 forecasts allow, only 10 throw up differences that are statistically significant at 95%. Given that, by construction, we would expect 5% of these tests to throw up significant results even if all the forecasts were equally good, I'm not sure we should read much into these results⁹.

Rare events

If you need time to distinguish two models when you have a regular flow of (noisy) data, this conclusion is only stronger – unsurprisingly – when you're trying to predict things that happen only occasionally (financial crises, for example).

Imagine now that we're trying to predict not a continuous variable, like GDP growth, but the occurrence (or not) of a discrete event that occurs randomly. And assume, for the moment, that there is nothing to know but the average frequency with which it happens. There are again two forecasters, each of whom has his, or her, own estimate. It's possible this time that neither is right; we, the observers, must decide which of them is better. What we're interested in is how long this might take.

Chart 9: Probability of rejecting λ_2 in favour of λ_1 , given true frequency λ



The lines in Chart 9 provide some examples. Specifically, for various values of the frequencies (the truth plus the two “models”) they plot the per-cent likelihood that we will be able to reject the less good model, at 95% significance, against the length of the sample. For example, suppose that one forecaster says the per-year probability of an event is 4% (we can expect it to happen once every 25 years) the other says it's 8% (once every 12½ years) but the true number is 12% (once almost every 8 years). This is the situation described by the blue line.

Given that one model is twice as far from the truth as the other, you might hope we could reject it in fairly short order. In fact, to have a better-than-evens chance of doing so, you'd need a sample of seventy years. You might be lucky, and see enough crises before then to reject the low-probability model. It would also take

⁹ We use a simple t-test of the difference in squared residuals (Diebold and Mariano (1995)). The comparison asks how likely it would be to get the differences we see in the sample if the two models in question were identical (the “null hypothesis”). If the answer is “less than 5%” we then conclude the two models are different. By construction, therefore, the probability of concluding that one model has outperformed another when both are, in fact, identical (a “Type I” error) is 5%.

a shorter time if you were less exacting about the test criterion. But, on these parameters, you'd need almost a lifetime of data to endorse the better model.

The reason is that all the frequencies involved – and above all the true frequency – are low. We simply don't get enough information, per unit of time, to make these distinctions, even when the two hypotheses we're testing are very different. In fact, if the second of the two forecasters were actually correct about the frequency – if it really was 8% – then it would take a good deal longer – 200 years, in fact – to be confident of rejecting the alternative in favour of the truth (the red line). Reduce these numbers only a little further (bad model 2%, true model 3%) and you'd need more than a thousand years of data to have even half a chance of telling one from the other. To all intents and purposes, you would never know.

I don't mean to pretend that this all models of financial crises (or any other discrete event) amount to. Real-world models of the financial system – and they are multiplying by the day – are aimed not at producing an estimate of the bare, unconditional likelihood of these events but at understanding what makes them more or less likely. Only then can we monitor the risks or have an idea how, at least cost, to reduce them¹⁰.

Nor do I want to sound too pessimistic about our ability to learn about these relationships. Financial crises may be infrequent, but they've happened in many countries, multiplying the amount of useful data. Nor are we limited to macro-economic data. For example, we may well learn useful information about the risks to the financial system from detailed analysis of individual banks' balance sheets.

But the point behind the simulations – that learning about infrequent events, and distinguishing between forecasts of those events, takes time – remains. Whether it's the simple frequency with which they occur, or a more sophisticated understanding of the risks their occurrence, we need quite a bit of data to uncover these things with any degree of precision. In the meantime, it probably won't be possible to make decisive comparisons between forecasters: you're likely to have to go through several events, and wait a long time, before deciding which is better.

Structural shifts and persistent errors

As Stockton points out, one of the features of the past few years, at least since 2010, is the persistence of forecast errors: growth has repeatedly turned out weaker than the MPC and others had expected, inflation higher.

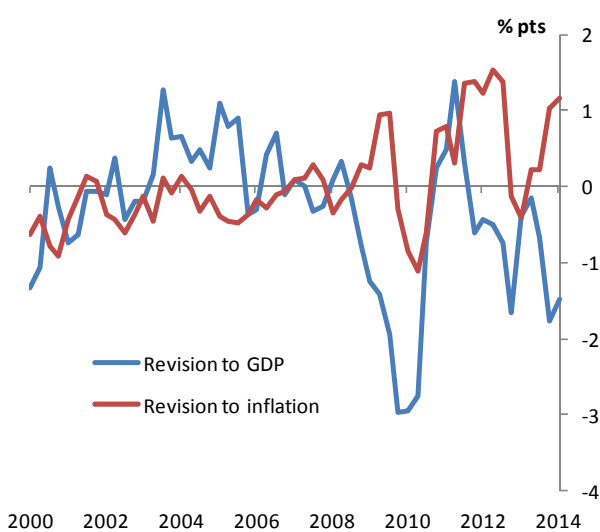
There are, he accepts, several identifiable shocks – things that could not fairly be described as predictable – that can account for some of this. Inflation has been boosted by persistent rises in commodity prices, the

¹⁰ The unconditional frequency of crises might still be a number worth knowing – it could affect the optimal baseline amount of capital banks should hold, for example (see Miles et al. (2011)). But if macro-prudential policy is to have anything to go on (if we want to know how to vary capital ratios, for example) we will need to understand what these risks depend on.

increase in the headline rate of VAT and successive increases in some administered and regulated prices. Higher commodity prices have also contributed to weaker economic activity; more importantly, output growth has also suffered from adverse developments in the euro area that were only partially anticipated by financial markets in 2010.

As we've learned, we should also be cautious about inferring too much from a relatively small sample. Even an unbiased model has a one-in-four chance of making errors in the same direction three times in a row, a probability well beyond the thresholds we normally view as statistically significant.

Chart 10: Revisions to MPC's two-year ahead forecast of growth and inflation¹



But the Inflation Report's longer-term growth forecasts have also been revised down more often than up in recent years (Chart 10 shows the changes in the forecasts for each date as it moves from two years to one year ahead). It's therefore reasonable to ask whether the MPC and other forecasters have taken too long to register the full impact of the financial crisis. The question is more easily asked than answered, however. The problem is that, after a significant and persistent shock, you may well need time to learn about its implications. As you do so, you are likely to make repeated forecast errors in the same direction.

¹Revisions calculated as the forecasts for period t made in period t-1 less that made in period t-2

Source: Bank of England

Let me use another simulation to get this point across. Suppose that, rather than having a single underlying, "trend" rate of growth, the economy can flit between two states, one "low-growth" state in which the local trend is zero, another in which it is 2%. Transitions between the two states are relatively infrequent: the model is set up so that, on average, the high growth states last six years, low growth states three years (this is meant, very roughly, to mimic the average duration of post-war expansions and downturns in the UK).

However, because there are also random, and moderately persistent¹¹, shocks to growth within each state, it's not clear straightaway, to an observer who gets to see only output itself, that a transition has occurred. For example, suppose the economy has been in its high-growth phase for a while and then suddenly weakens. This could be because it has switched state, in which case growth is likely to remain low for a

¹¹ The first-order autoregressive coefficient for annual GDP growth in the UK, over 100 years, is around one third: if growth is 1% point higher than average one year, the best guess is that it will be around 0.3% points higher the following year. This is the coefficient we use for the within-state shocks in this simulation.

while. But it could be because, within the high-growth state, we happened to experience a bad but passing shock, in which case we should expect growth to resume. Not knowing for sure which of these is true, the best possible forecast will give some weight to both possibilities. After a genuine switch, the observer will therefore over-predict growth. And he or she will continue to do so until it becomes clear the transition has occurred. The red line in Chart 11 plots the expected path of forecast errors after a transition from the high to the low-growth state.

Again, I do not mean this as a realistic description of how forecasts of growth are constructed, still less of the economy itself. But I think it captures an important point, namely that if you have to learn about the economy as you go along, you are more likely to make serially correlated forecast errors. This looks like bias after the event. But, given the information available at the time, it is not. Indeed, it's only through making these repeated errors that the forecaster is able to learn that a shift has occurred.

And the more there is to learn, the larger this effect is likely to be. For example, suppose that, at a certain point it's not just the state of the economy but its structure too that changes: the average duration of the low-growth state jumps from three years to seven years (even more roughly, this is intended to capture something about the extended weakness that tends to follow financial crises, as opposed to "normal" recessions). The forecaster doesn't know this, however: he or she has to work it out over time. And at least in this simulation, it takes several years to do so (Chart 12 plots the best possible estimate of the duration of the low-growth state to an observer in the simulated economy). As a result, the size and the persistence of forecast errors are, in the meantime, likely to be all the greater (the blue line in Chart 11).

Chart 11: Forecast errors are serially correlated after a structural break

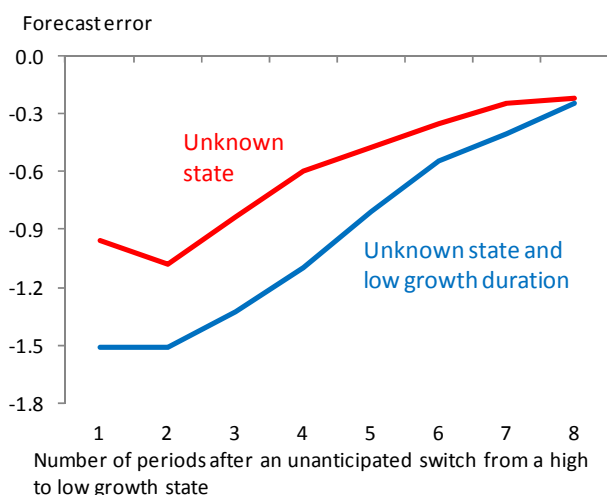
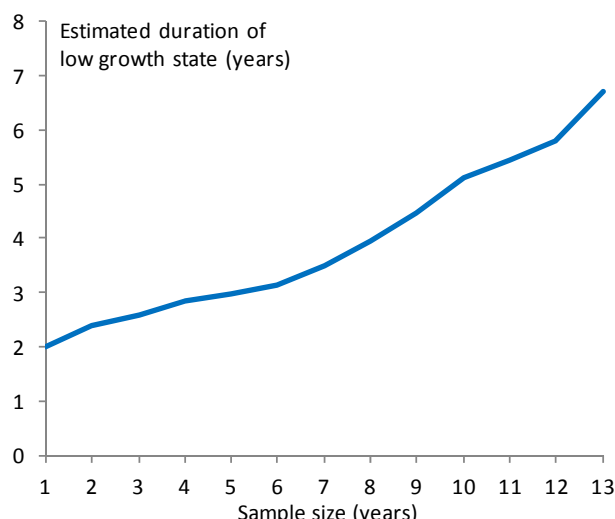


Chart 12: Learning about a change in a model parameter takes time



I have deliberately set up these examples to make a point. And you could easily dispute their realism, in particular the key assumption that the forecaster has only the history of growth itself to go on. In 2009 we

knew there'd been a financial crisis and we also knew that, following similar events in the past, economic growth had been weaker than after other recessions (Reinhart and Rogoff (2009)). So forecasters should not excuse themselves by claiming that the most reasonable assumption in 2009 was that the economy would follow the path of recoveries after typical (non-crisis) recessions.

I think there may be something to this, at least in my case. In my last job I too over-predicted GDP growth in 2011 and 2012, significantly so. And although I can blame some of that on unpredictable events, I also have a sense that I failed to attach enough weight to the historical experience. Implicitly, I suspect, I was assuming a more rapid cleansing of the balance sheets of banks, and a faster improvement in the general functioning of the financial system, than had occurred after other financial crises, but with no obvious reason to do so.

But the point remains that, inevitably, we are to some extent having to learn about the economic implications of the crisis as we go along. There may have been others, but large financial crises are rare events and none is exactly like any other. So I think it's legitimate to view this as a "structural break" – a shift in things that, for a long period of time, we've been happy to treat as unchanging constants (e.g. trend growth) – and something we can understand only over time¹². If so, even the best forecast will tend to be wrong in the same direction, at least for a while.

Statistical inference and human inference

The last section pointed out that, the greater the degree of randomness in a series, the longer the sample you need to judge forecast performance. That sample length was all the greater, unsurprisingly, the rarer the event you're trying to forecast. And when the rare event is a shift in the entire model we've been using, we are likely to make repeated errors in the same direction. In one form or another, therefore, I've merely been pointing out that the world is unpredictable and that you may need a lot of data to distinguish true structure from what is just chance.

This is surely uncontroversial. You may even think that it's blindingly obvious. But we often seem to behave in ways that suggest we do not, intuitively, grasp the point. In his book "Thinking Fast and Slow", the psychologist Daniel Kahneman describes eloquently how prone we are to under-weighting randomness in small samples. If you toss a coin four times in a row you are much more likely to get a sequence of uninterrupted heads (or tails) than if you do so seven times. Yet someone drawing the smaller sample is just

¹² One thing whose behaviour differs radically from that in previous cycles, to an extent consistent with the idea of a "structural break", is UK productivity. As I pointed out in a speech last year, the chance that, given the path of output, and given the previous relationship between the two, employment would turn out as strong as it has been since 2008, is around one in five hundred. This departure from prior norms is reflected in the continuous and almost universal over-prediction of productivity growth in the past five years. In its annual survey of UK economist forecasts, the Treasury has published a total of over 70 separate year-ahead productivity forecasts since 2008 (around 14 per year). Only three of these – all for 2010, when the economy turned out stronger than many had expected – under-predicted productivity growth. The rest were all too high. Rather than viewing this as a shared and correctable bias, I think a more reasonable interpretation is that, for whatever reason, and for however long, the behaviour of the economy has shifted, and in a manner that – even if we succeed in explaining it after the event – was not knowable in advance.

as likely to interpret these sequences as evidence of bias¹³. Sports fans are too readily convinced that an individual player's sequence of successes – sequences which, given enough random variability in performance, are bound to occur from time to time – are evidence of “good form” that will persist into the future. The finance industry rewards people for predictive success when that is often just luck, and unlikely to endure¹⁴.

It's not clear why evolution has made us like this. But, wherever it comes from, the tendency to see the world as more deterministic than it really is has clear knock-on effects. It means, for example, that we systematically under-estimate the likely error in our own judgements. Faced with general knowledge questionnaires that ask not just for quantitative answers but the confidence bands around them, people repeatedly make those bands too narrow: the true answer lies outside the intervals far more often than people anticipate¹⁵. In similar experiments people who've reported being “100% certain” of something turn out, on average, to have been correct on 70%-80% of occasions¹⁶. And this “over-confidence effect” applies to professionals as well as amateurs. One of the first experiments to establish it involved psychiatrists themselves (Okamp (1965)). More recently, the political scientists questioned by Philip Tetlock had little idea that their predictions, at least in this instance, performed no better than purely random guesses¹⁷.

A related failure is “hindsight bias”, the tendency to see events that have already occurred as being more predictable than they really were. Like over-confidence, evidence for this bias is widespread and well-established¹⁸. People are prone retrospectively to exaggerate the probability they'd assigned to events that did occur, and to understate the likelihood they'd attached to things that did not. And the interesting thing is that this tendency is not just about appearances: unless reminded with hard evidence, people seem genuinely to believe that their prior predictions were different from what they actually were. The tendency to absolve ourselves of past predictive error is therefore deep-seated and, unless consciously checked, automatic. And, as Kahneman points out, it can have “pernicious effects on the evaluations of decision makers...we are prone to blame [them] for good decisions that worked out badly and to give them too little credit for successful moves that appear obvious only after the fact”.

To a new reader this literature is fascinating but humbling. One can be forgiven for worrying that we human beings are nothing but a collection of neuroses and self-regarding biases¹⁹. Nor, as I say, are (so-called) professionals remotely immune from these failures. But all is not lost: we are creatures of reason as well as

¹³ The converse is also true: when asked in an experiment to generate random sequences (of coin tosses, for example), people generally include too much alternation and too few sequences of the same outcome (Bakan (1960)). A more recent example didn't require an experiment. In 2010, Apple had to alter the “shuffle” function in its iPod to make more switches of tracks than a truly random sequence would generate, as users had complained (wrongly) that the original function contained too few to be properly random.

¹⁴ Taleb (2001). This is not to say that these reactions are always wrong. It may be that “form” – persistent outperformance, whether in sport or finance – does exist. It's just that, when faced with a sequence of good results, we are too willing to attribute it to something persistent and too unwilling to allow for the possibility that it's random.

¹⁵ A survey article in 1982 found that, across a number of studies asking for 98% confidence intervals (around what came to a total of 15,000 questions) the true answer lay outside these bands 32% of the time (Lichtenstein et al (1982)).

¹⁶ Fischhoff et al (1977).

¹⁷ Tetlock (2005).

¹⁸ See, for example, Blank et al. (2007).

¹⁹ The Wikipedia page on the subject lists 171 different recorded cognitive biases.

unthinking instinct (otherwise we wouldn't even be able to recognise these biases). And, applying that reason, we can counter some of our natural weaknesses. For example, one experiment found that, if its subjects were required to add to their general knowledge answers a list of reasons why they might be wrong, the overconfidence effect was much less marked²⁰.

For the same reason, economic forecasters, including the MPC, should continually expose themselves to question – as Stockton recommends – and keep in mind that, over the past, we've been able to predict only a minority of movements in GDP growth, even from only a year away. Clearly, the same obligation should apply to those who judge forecasts.

Conclusion

Recently I spoke to someone who told me that, ahead of the financial crisis, he'd seen a model that predicted its occurrence "with 100% certainty". This struck me as odd. It's not just that, death and taxes apart, nothing can be predicted with 100% certainty. Nor is it simply that, if there had been such a model, there probably wouldn't have been a crisis: what were all those people doing buying and financing assets if their (the assets') demise had been as inevitable as an apple falling from a tree? What was striking is that, after an event that one would have thought should make us less certain about the world, he had become more certain: his beliefs about the causes, the prior likelihood and the consequences of the financial crisis were settled and definitive. And that meant he also viewed the failure to predict the event, or the weak growth that has followed, as genuine and avoidable mistakes.

To some degree that may well be true. Clearly, it would be unforgivably complacent not to learn from this experience. As Stockton says, "the observation that forecasting errors were widespread during this period does not obviate serious introspection on the part of economic forecasters about what went wrong and what lessons might be learned". And that introspection could well lead to the realisation that there were avoidable mistakes.

At the same time, we should remember that it is only through forecast errors – by coming across things we hadn't previously thought of – that we discover more about the world. "We should be pleased with forecast failures," says Sir David Hendry, the distinguished econometrician, "as we learn greatly from them". Yet, in reality, we do not always find it a pleasing experience. We all of us prefer to be right and are made uncomfortable by events that don't fit into a coherent model of the world, preferably the one we already hold in our heads. Psychologists tell us that these instincts are so deep-seated that they often over-ride our rationality: we wishfully see structure in random events; believing this structure, we are often over-confident about our own predictions; when it comes to others', we are too quick to assign significance to their forecasting errors, whether small or large. If the forecast turns out to have been correct we immediately

²⁰ Hoch (1985).

assume the forecaster is good; when it's wrong we are quick to blame the forecaster rather than chance. As we saw with some of the simulations it can, with enough chance, take a very long time to tell apart a "good" from a "bad" forecast, but our instincts often jump the gun.

I think these pitfalls are sometimes apparent in the coverage of economic forecasts, including the MPC's. In a speech in 1999, the Governor said it should be easier for central banks than it is for governments to admit our ignorance about the future:

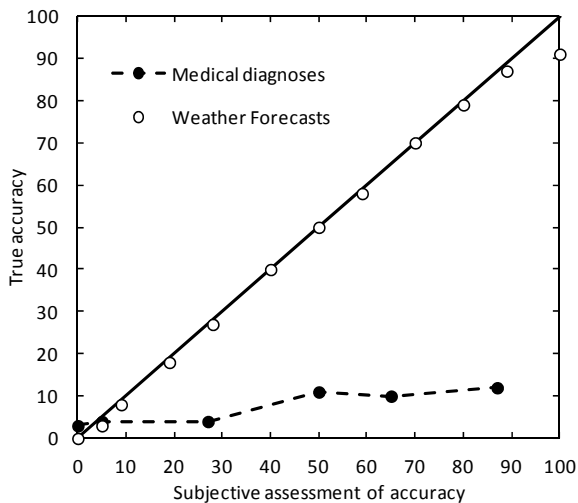
"Perhaps one of the strongest arguments for delegating decisions on interest rates to an independent central bank is that, whereas democratically elected politicians do not often receive praise when they say "I don't know", those words should be ever present on the tongues of central bankers."

It was with this in mind that, when the Inflation Report was first launched, the projections of growth and inflation were represented not as single numbers but as a distribution of possible outcomes – the "fancharts". Yet the Inflation Report forecasts are still almost always reported as precise, point predictions. When, in an Inflation Report press conference in 2011, the Governor asked rhetorically "Who knows what's going to happen tomorrow, let alone in the next twelve months?" there seemed to be consternation, not to say alarm, that he should not, in fact, have a firm idea of the future. And although Stockton himself never reached this conclusion – he merely said that subsequent events "may not fully explain" the size or persistence of the Bank's forecast errors, and focused most of his attention on the forecast process, not its outcome – the coverage of his report was overwhelmingly about the forecast errors and the failure that they represented.

I don't want to belabour these points. I should certainly not leave you with the impression that economic forecasting is so inaccurate that we shouldn't bother with it. For one thing, we have to: central banks cannot avoid making judgements about future risks, in some form or other, because monetary policy only works with a lag. As Alan Greenspan put it some years ago, "Implicit in any monetary policy action or inaction is an expectation of how the future will unfold, that is, a forecast" (see also Budd (1998)). Second, the usefulness of the Inflation Report process extends well beyond the production of the fancharts: it facilitates a detailed discussion of the implications of alternative policies and allows the MPC to communicate its views to the public.

Nor should I be too defensive. I've tried to make the point that you may need a lot of data before valid criticism of the performance of economic forecasts becomes possible. But we should always be interested in how to improve them and alert to the possibility that they are, in fact, flawed. Besides, even invalid criticisms have their uses.

Chart 13: Actual versus predicted accuracy in two surveys of professional judgement



Source: Plous (1993)

Psychologists have discovered that not every profession suffers from habitual overconfidence. Chart 13, for example, taken from a book by the psychologist Scott Plous, reveals that – at least according to the two surveys he quotes – weather forecasters are much more realistic than doctors about the accuracy of their judgements²¹. He attributes this to the continual reminders that weather forecasters receive about how inaccurate they can be. So, while it may wrongly presuppose the existence of some much better forecasting process, one that doesn't make “errors”, the criticism of the MPC's predictions should at least keep us honest about our limitations.

²¹ The judgements are projections of precipitation, at various time horizons, in the case of the weather forecasters, pneumonia diagnoses in the case of the doctors. In each case, the horizontal axis represents the subjective assessment of accuracy (for example, the weather forecasters' confidence about their own projections so many days ahead), the vertical axis the actual accuracy. The further below (above) the diagonal, the greater the degree of unwarranted over(under)confidence.

References

- Bakan, P. (1960)**, "Response tendencies in attempts to generate random binary series", *American Journal of Psychology*, 73, 127-131
- Bank of England, 2013**, "Response of the Bank of England to the three Court-commissioned reviews". Bank of England, March 2013.
http://www.bankofengland.co.uk/publications/Documents/news/2013/nr051_courtreviews.pdf
- Budd, A., 1998**, "Economic policy, with and without forecasts". Speech given at the Sir Alec Cairncross Lecture for the Institute of Contemporary British History and the St. Peter's College Foundation, 27 October 1998, available at:
<http://www.bankofengland.co.uk/publications/Documents/speeches/1998/speech28.pdf>
- Dale, S., 2011**, "MPC in the dock". Speech given at the National Asset-Liability Management Global Conference, London, available at:
<http://www.bankofengland.co.uk/publications/Documents/speeches/2011/speech485.pdf>
- Diebold, F.X. and Mariano, R.S., 1995**, "Comparing Predictive Accuracy". *Journal of Business and Economic Statistics*, 13:3, 253-63.
- Elder, R., Kapetanios, G., Taylor, T. and Yates, T., 2005**, "Assessing the MPC's fan charts". *Bank of England Quarterly Bulletin*, Autumn 2005.
- Fischhoff, B., Slovic, P., and Lichtenstein, S., 1977**, "Knowing with certainty: The appropriateness of extreme confidence". *Journal of Experimental Psychology: Human Perception and Performance*, 3, 552-564.
- Greenspan, A., 1994**, "Discussion". In Forrest Capie, Charles Goodhart, Stanley Fischer and Norbert Schnadt. *The Future of Central Banking*, Cambridge University Press.
- Hendry, D.F. and Ericsson, N.R., 2001**, *Understanding Economic Forecasts*. MIT Press.
- Hoch, 2005**, "Counterfactual reasoning and accuracy in predicting personal events". *Journal of Experimental Psychology*, 11, 719-731.
- Kahneman, D., 2011**, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Kim, C-J. and Nelson, C.R., 1999**, *State-Space Models with Regime Switching*. MIT Press.
- King, M., 1999**, "Challenges for Monetary Policy: new and old", *Bank of England Quarterly Bulletin*, 39(4).
- Lichtenstein, S., Fischhoff, B. and Phillips, L.D., 1982**, "Calibration of probabilities: The state of the art to 1980". In Daniel Kahneman, Paul Slovic, Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. pp. 306–334.
- Miles, D., Yang, J. and Marcheggiano, G., 2011**, "Optimal bank capital". External MPC Unit Discussion Paper No.31: revised and expanded version.
<http://www.bankofengland.co.uk/publications/Documents/externalmpcpapers/extmpcpaper0031.pdf>
- Oskamp, S., 1965**, "Overconfidence in case-study judgements". *The Journal of Consulting Psychology (American Psychological Association)* 2: 261–265. reprinted in Kahneman, Daniel; Paul Slovic, Amos Tversky (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. pp. 287–293.
- Plous, S., 1993**, *The Psychology of Judgment and Decision Making*. McGraw Hill.

Reinhart, C.M. and Rogoff, K.S., 2009, *This Time is Different: Eight Centuries of Financial Folly*. Princeton University Press.

Rumsfeld, D., 2002, “Transcript of DoD news briefing – Secretary Rumsfeld and Gen. Myers”. US Department of Defence, February 12 2002.
<http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636>.

Stockton, D., 2012, “A review of the Monetary Policy Committee’s forecasting capability”. Report by David Stockton presented to the Court of the Bank of England, October 2012.
<http://www.bankofengland.co.uk/publications/Documents/news/2012/cr3stockton.pdf>

Tetlock, P., 2005, *Expert Political Judgment: How good is it? How can we know?*. Princeton University Press.

Taleb, N., 2001, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Random House, New York.