

Do our forecasts improve as we finalise them?

Sharon McCaw

Editor's note

The motivation behind this paper was to examine whether the economic judgement added to our forecasts improves them, versus an 'unadulterated' FPS-model run. However, we quickly realised that we did not have the data to enable us to examine this question precisely. Although we have archived earlier versions of our forecasts, it is not easy to separate out the different influences on these forecasts. The main issues are:

- New data is always coming to hand as we finalise the forecasts over several weeks – some 'judgement' reflects this.
- Particularly in the early period in which we first started using FPS, some 'judgement' was not economic, but rather technical fixes.
- The FPS model itself has gradually been recalibrated to better reflect policy-makers' views of the economy. Judgement is therefore embedded with the model structure.

Given these issues, we could not truly analyse the role of judgement, but rather only ask the less interesting question, "do our forecasts improve as we finalise them?"

In brief

This paper examines whether our forecasts improve in terms of accuracy (size of errors) and bias (mean error) during the process of finalising our forecasts using the Forecasting and Policy System (FPS) model. We investigate the marginal impact of judgement on the accuracy and bias of our forecasts of inflation, the output gap, interest rates and GDP growth since 1997.

Neither the accuracy nor bias of the projections was statistically significantly different between any of the forecast iterations, for any variable, at any horizon. This is mainly due to a paucity of data. Any conclusions are therefore highly tentative.

It appears that as the forecasts progress we have tended to lower our output gap and interest rate forecasts on average, but analysis reveals that four particular quarters dominate this result. Two of these were during the Asian crisis, when the world was changing rapidly. Of the other two, one took place in the early days of the FPS model when we were still adjusting how we would measure potential output, and one in 1999 when GDP growth signals and direct output gap indicators were giving quite different signals.

Introduction

Since 1997 we have been using the Forecasting and Policy System (FPS) model to generate our macroeconomic forecasts. As we gradually refine our forecasts over a period of several weeks, we input more recent data but also impose judgements on the model regarding likely developments in the economy. The FPS model ensures that the forecasts remain internally consistent, but is quite flexible in terms of allowing judgement to be imposed on the model in a non-mechanistic manner. In addition, the preliminary versions of the forecasts are able to be archived.

This paper examines how our forecast performance for key macroeconomic variables changes as forecast rounds progress. The following rounds were identified for forecasts from June 1997 to March 2002:

- A ‘first cut,’ our first ‘sensible’ run after basic technical issues have been ironed out and our first estimate of the near-term situation has been decided upon.
- An ‘MPC’ run presented to the Monetary Policy Committee (MPC) for their consideration.
- The ‘final’ (published) run after the MPC have added their judgement.

Not all the changes between runs represent ‘judgement’ in terms of off-model views on the current economic situation; additional economic data also continues to come to hand as the forecast rounds progress over several weeks. This can be quite influential in volatile times, such as when the Asian crisis was unfolding. It is infeasible to separate out these effects after the event. A further caveat when interpreting the findings is that the FPS model has been gradually recalibrated to include much of the thinking that entered the early rounds as ‘judgement’. There is therefore no time-consistent interpretation of what the judgement being added represents, as the model to which that judgement was being applied was in a constant state of change.

Previous Reserve Bank analysis, though it had a very short sample, concluded that judgement had possibly made our annual inflation forecasts more accurate.¹ Our sample size is still poor: it ranges from only 11 observations for the 10-step ahead forecasts, to 20 observations for the 1-step ahead forecasts. Neither the accuracy nor bias of the projections was statistically significantly different between any of the forecast iterations, for any variable, at any horizon. Any conclusions are therefore highly tentative.

This paper focuses on annualised quarterly CPI inflation,² 90-day rates, the output gap and quarterly GDP growth. It examines the impact of judgement on the size of our forecast errors (the ‘accuracy’), and also, unlike previous analysis, the bias.

1 CPI forecasts

We first examine the variable subject to our most clearly biased forecasts: CPI inflation. Note, however, that in the FPS model, the forward path of interest rates always adjusts to ensure that projected CPI inflation outcomes are consistent with the Policy Targets Agreement. As a result, although we sometimes add judgement to inflation directly, judgement added to the forecasts will, in the medium term, generally impact much more on the interest rate and the output gap forecasts than on the CPI projection.

1.1 Accuracy

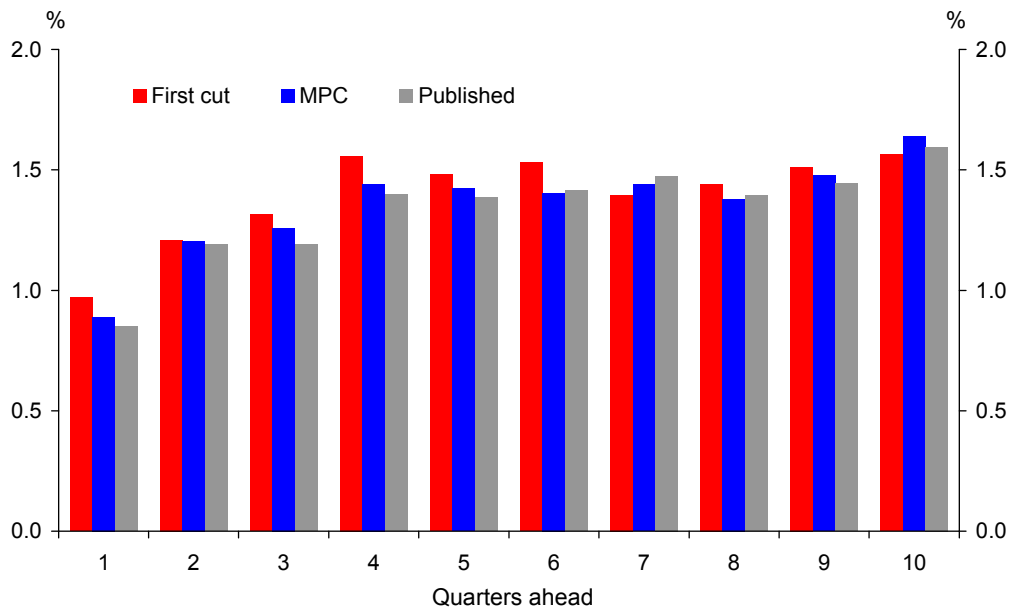
Figure 1 below gives the mean absolute error (MAE) for the three sets of our FPS annualised quarterly CPI inflation forecasts.³

¹ St Clair, R (2000), “The impact of judgement on forecasting performance: some preliminary insights.” Reserve Bank of NZ *Memorandum*

² This aims to avoid the serial correlation problems inherent in looking at annual CPI errors. Using annualised data makes interpreting the size of the errors more intuitive; it has no impact on results.

³ Given the small sample sizes, outliers strongly dominated the root mean squared error statistics. We therefore focus on the MAE. RMSE results are available on request.

Figure 1
CPI - Mean absolute error



There are few clear patterns, though around one-year ahead it appears that the accuracy of the forecasts may improve slightly as we iterate.

1.2 Bias

Figure 2 shows the mean errors (‘bias’ when statistically significantly different from zero) of the CPI forecasts from the various runs.

Figure 2
CPI Mean forecast errors

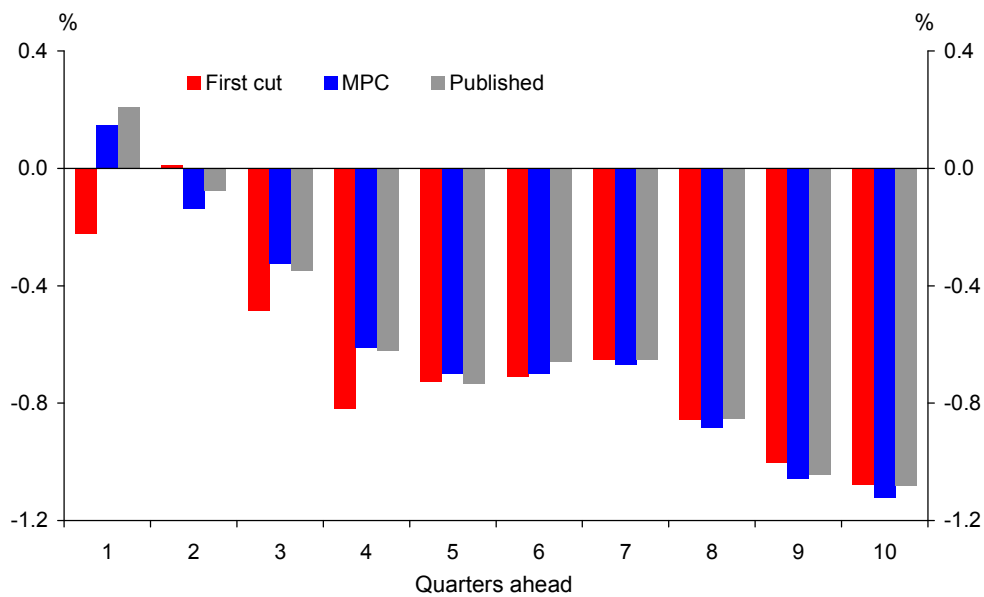


Table 1 below summarises the statistical significance of the mean errors graphed above, ie whether there is a demonstrable bias.⁴

Table 1
Statistical significance of mean error (bias) in annualised quarterly CPI errors

Steps ahead:	1	2	3	4	5	6	7	8	9	10
First cut	-0.23	0.01	-0.48	-0.81*	-0.72	-0.69	-0.64	-0.84*	-0.98*	-1.05**
MPC	0.15	-0.14	-0.32	-0.61	-0.70	-0.70	-0.67	-0.89*	-1.06**	-1.12**
Published	0.21	-0.08	-0.35	-0.62	-0.73*	-0.66	-0.65	-0.85**	-1.04**	-1.08**

Notes to the table:

For the CPI '1 step ahead' is for the calendar quarter in which the forecasts are produced.

* = Significantly different from zero at 10 per cent level of significance

** = Significantly different from zero at 5 per cent level of significance

*** = Significantly different from zero at 1 per cent level of significance

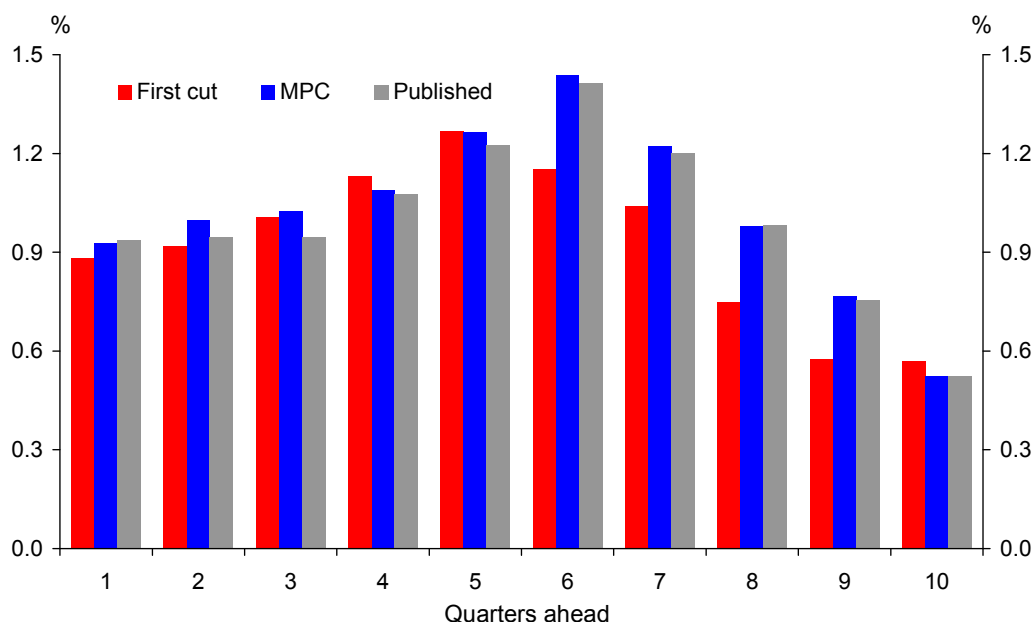
There is no consistent pattern regarding whether judgement improves or worsens the bias in our inflation forecasts, and as noted, no statistically significant differences.

2 Output gap forecasts

2.1 Accuracy

Figure 3 shows that between 6 and 9 steps ahead the judgement added after the first cut, before the forecasts are presented to MPC, has tended to make our gap forecasts less accurate. However, once again the differences were too small to be statistically significant.⁵

Figure 3
Output gap – mean absolute error



⁴ Note: the reason that these results are statistically less significant than those reported in earlier analysis is the smaller sample period (as we can use only the post-FPS period for this analysis). However, the bias pattern is consistent. Also note that these are for annualised quarterly inflation and must be divided by four for approximate comparability with figures given in other papers.

⁵ Note that the sample size also shrinks as the forecast horizon increases; at 10 steps ahead we have only 11 observations, and the key statistics are therefore imprecisely estimated.

2.2 Bias

On average we have tended to lower the forecast output gap as the forecasts are finalised (figure 4) (though again, none of the mean errors are statistically significantly different from each other.) In all runs the bias is strongest in the very near-term and reduces (and becomes insignificant) further out.

Figure 4
Output gap mean error

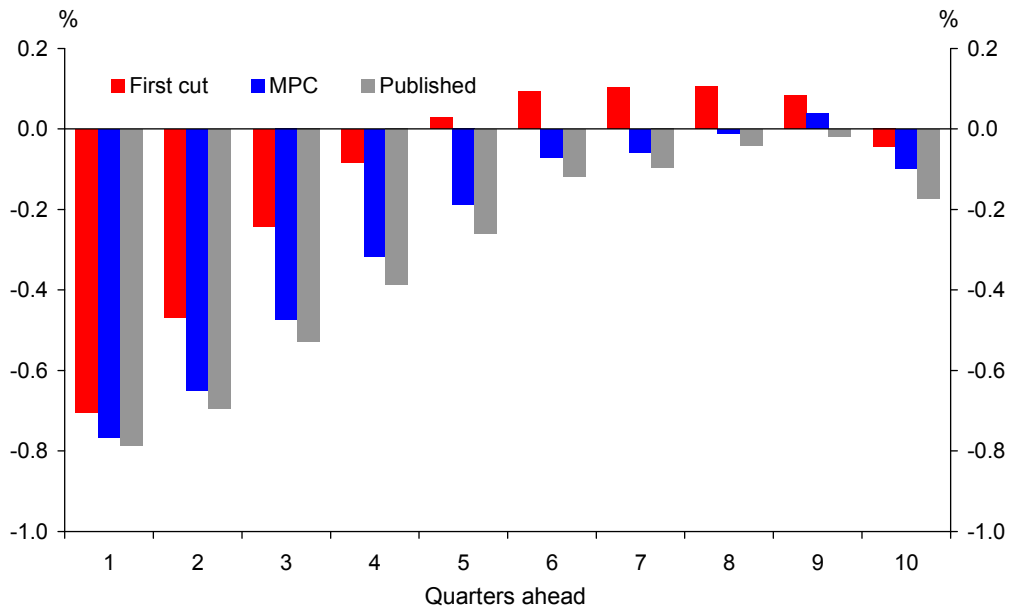


Table 2
Statistical significance of mean error (bias) in output gap forecast errors

Steps ahead:	1	2	3	4	5	6	7	8	9	10
First cut	-0.7***	-0.47	-0.24	-0.08	0.03	0.09	0.10	0.11	0.08	-0.05
MPC	-0.77***	-0.65**	-0.47	-0.32	-0.19	-0.07	-0.06	-0.01	0.04	-0.10
Published	-0.79***	-0.69**	-0.53	-0.39	-0.26	-0.12	-0.10	-0.04	-0.02	-0.17

Note:

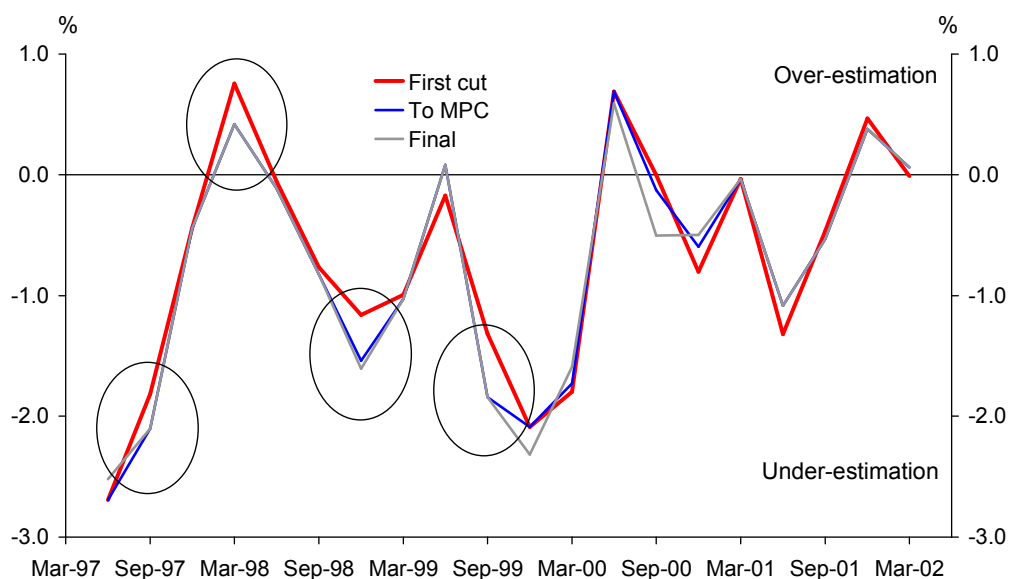
- * = Significantly different from zero at 10 per cent level of significance
- ** = Significantly different from zero at 5 per cent level of significance
- *** = Significantly different from zero at 1 per cent level of significance

The starting point output gap estimates and one quarter ahead forecasts errors are biased towards under-prediction by at least 0.6 percentage points, and are therefore potentially an important source of inflation bias. It is worth noting that this bias exists in all of the runs. However, the output gap starting point under-estimation bias closes more quickly in the first cut projections than in the later runs.

However, analysis of the changes in the forecasts through time reveals that the largest adjustments to the output gap estimate, which dominate these results, were in four quarters: September 1997, March 1998, December 1998, and September 1999 (figure 5). During these quarters we revised down our output gap estimate by about half a per cent. In March 1998 this improved our estimate, but in the other three quarters it worsened the under-estimation.

Aside from these four quarters, the output gap errors have been smaller and more evenly distributed between under- and over-estimation.

Figure 5
1-step ahead output gap level forecast errors



Examining these four quarters more closely reveals that in March and December 1998 we revised down our estimates of growth considerably, in the first case improving the forecasts, but in the latter worsening them. The motivation in both quarters was the unfolding Asian crisis and concerns about the outlook for our trading partners. These findings are consistent with our conclusion in the Business Cycle Review (1999) that we were a little slow in recognising the seriousness of the Asian crisis, but then also under-estimated the speed at which the economy would come out of it.

In September 1997 and September 1999 we added judgement in order to revise up our estimate of potential output growth considerably. In September 1997 this reflected a change in our methodology for calculating potential output, putting more emphasis on projections of the capital stock and growth in total factor productivity. The September 1999 adjustment was a judgement call based on independent indicators of resource pressures. This improved our potential output estimates, but, by worsening our under-prediction of GDP, worsened our output gap estimates.

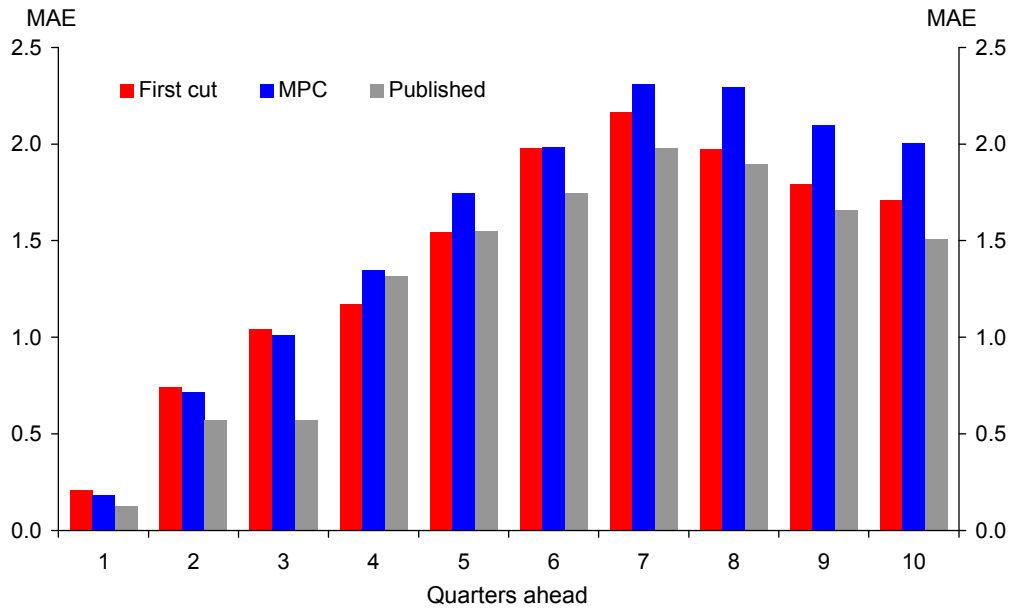
3 90-day rate forecasts

3.1 Accuracy

Figure 6 plots the MAE for our 90-day rate forecasts. 90-day rates are heavily influenced by our OCR decisions but also reflect market expectations of the OCR, which often differ from our published track.⁶ However, not surprisingly, given our influence over this variable, our final published forecasts are the most accurate predictors of actual rates. (Once again, however, the size of the errors is not statistically significantly different across the different runs).

⁶ Note also that a further caveat to the results is that until March 2001 the yield gap, rather than the level of 90-day rates, was the key variable influencing activity in the FPS model.

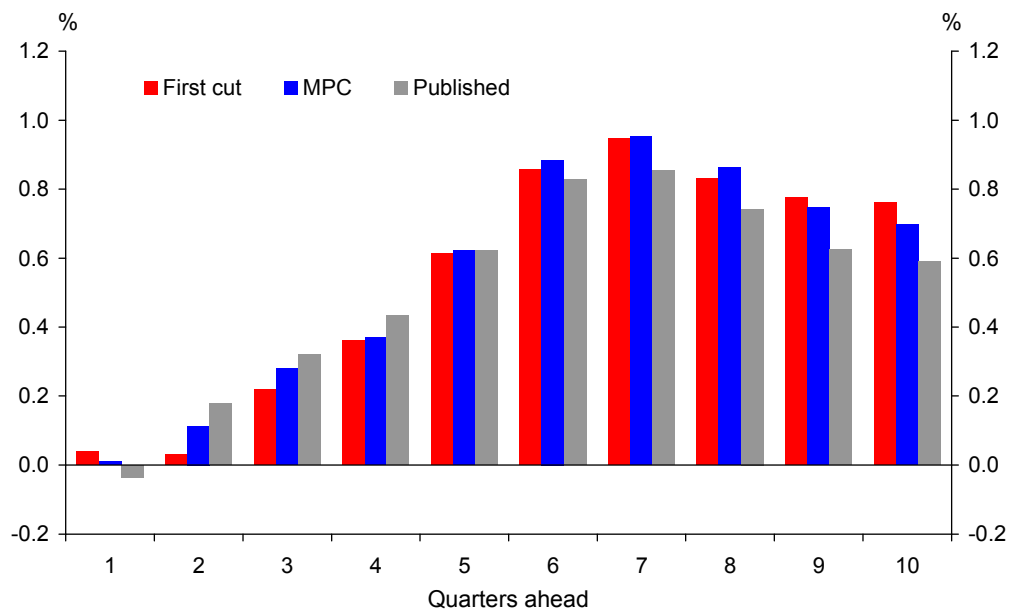
Figure 6
90-day rates – Mean absolute error



3.2 Bias

Our published forward 90-day rate tracks have, on average, been lowered as the forecasts are finalised, and thereby have possibly become a little less biased towards over-prediction beyond 6 quarters out (figure 7). (But note that the bias is not statistically significantly different from zero in any of the runs, nor are the runs statistically significantly different from each other).

Figure 7
90-day rates – Mean ‘forecast’ error



It may appear odd that the judgement applied to the forecasts two, three and four quarters ahead has served to increase the rate track on average, given that judgement has tended to lower our output gap forecasts. However, this is due to an outlier in the March 1998 quarter.

At this time we were heading into the Asian crisis. Forecasts of the TWI were being revised down and interest rate forecasts were accordingly increasing (although we were simultaneously revising down our forecasts of the world output gap).

Table 4
March 1998 exchange rate and interest rate forecasts

Forecast	TWI				90-day rates			
	1-step (98q1)	2-step (98q2)	3-step (98q3)	4-step (98q4)	1-step (98q1)	2-step (98q2)	3-step (98q3)	4-step (98q4)
MPC (20 Feb)	61.5	61.1	61.1	61.2	8.9	8.1	7.5	7.0
Final (27 Feb)	61.5	60.2	59.6	59.5	8.9	9.3	9.0	8.9
Actual	61.2	58.5	57.1	56	9.0	9.1	6.8	4.5

If this quarter is excluded from the analysis, the judgement applied to our interest rate tracks at the early horizons has in fact been towards lowering the interest rate tracks on average, as one would expect.

3.3 Putting it together

Although the differences are not statistically significant, the evidence tentatively suggests that as our forecasts have been finalised, there is a greater tendency for our output gap forecasts to be too low, and for our projected interest rate tracks to be lowered (in the process making them closer to the actual path of rates). However, this finding is focused in four specific quarters. Was the applied judgement appropriate? We conclude that in two quarters it was, while in two other quarters it was unhelpful. In other quarters judgement has not had a particularly large impact on findings.

Conclusion

- The bias in our estimate of the starting point of the output gap is important, and exists even in very early iterations of our forecasts. However, on four occasions, judgement has lowered our estimates considerably. Half of the time this has made the forecast more accurate, while the other two times it was unhelpful.
- Because on average over the period the output gap estimate has been revised down, the interest rate profile has on average been reduced also.
- Caveats to these findings include the short sample period, and the difficulty of differentiating the impacts of ‘judgement’, incoming data and model recalibrations.
- It is worth stressing again that there were no significant differences between any of the runs; the measured differences may therefore be very inaccurate representations of ‘typical’ behaviour.

Appendix 1: Production GDP forecasts

We also examined how our quarterly GDP growth forecasts have evolved during forecast rounds.

There is no consistent pattern in how the accuracy (MAE) of the forecasts changes as the forecasts are finalised (figure A1:1).

Figure A1:1
Quarterly GDP - Mean absolute error

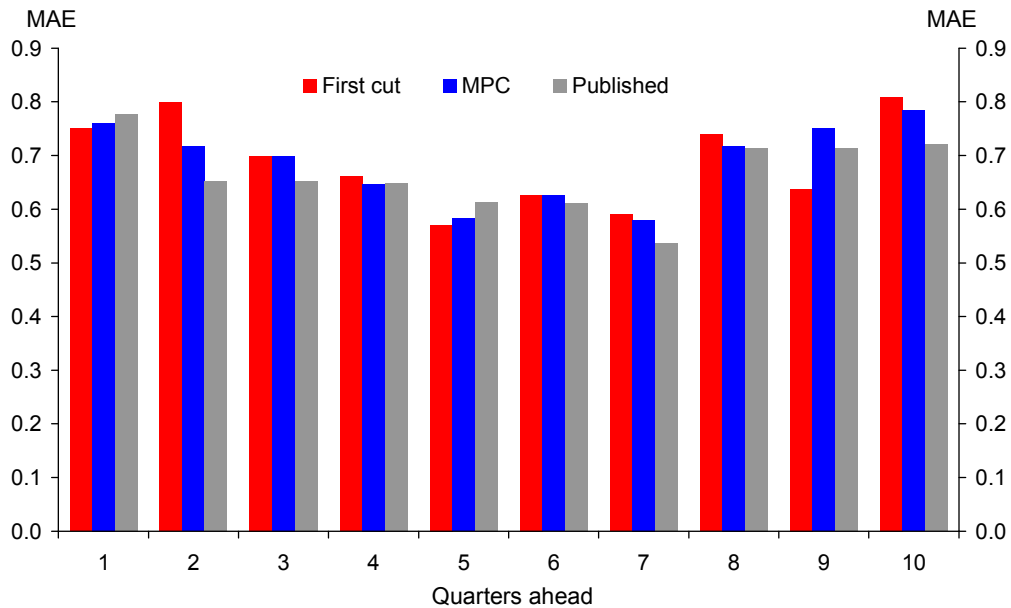
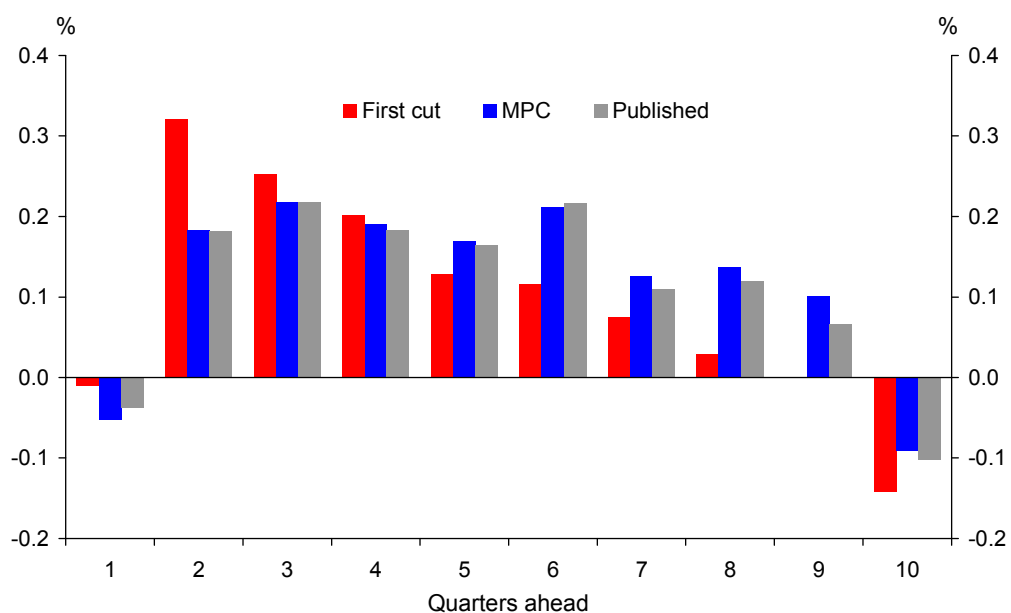


Figure A1:2 shows the estimated mean error ('bias' when statistically significant) of the quarterly GDP forecasts.

Figure A1:2
Quarterly GDP forecasts – mean errors



The mean error becomes less positive (less over-prediction) at longer horizons. Note that none of the mean errors are statistically different from zero.⁷

The bias was not significantly different between the forecast iterations. However, note that, although it is not significant, the second monitoring quarter GDP growth estimate has tended to become less biased towards over-prediction as the forecast rounds progress. This causes the disparity in the bias in our estimates of the output gap seen in figure 4 between one and two-steps ahead; the first cut eliminates the starting point bias towards under-estimation more quickly than the other runs because it has a significantly higher GDP growth estimate for the second monitoring quarter.

Note, however, that as discussed in the main body of the paper, these findings are dominated by a few particular quarters, largely falling in the relatively early days of the FPS model. In recent years the judgement has not been consistently in either direction.

⁷

Note that this is using 'latest' GDP data. Ranchhod ("[GDP forecast errors](#)") found significant bias when using the data as it stood a year after its initial release.