Quast, Josefine; Wolters, Maik H.

Article — Published Version

The Federal Reserve's output gap: The unreliability of real-time reliability tests

Journal of Applied Econometrics

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Quast, Josefine; Wolters, Maik H. (2023) : The Federal Reserve's output gap: The unreliability of real-time reliability tests, Journal of Applied Econometrics, ISSN 1099-1255, Wiley, Hoboken, NJ, Vol. 38, Iss. 7, pp. 1101-1111, https://doi.org/10.1002/jae.3003

This Version is available at:
https://hdl.handle.net/10419/288194

Received: 18 November 2021 | Revised: 19 January 2023 | Accepted: 21 January 2023

DOI: 10.1002/jae.3003

WILEY Journal of APPLIED ECONOMETRICS

REPLICATION

The Federal Reserve's output gap: The unreliability of real-time reliability tests

Josefine Quast[1] | **Maik H. Wolters**[2,3,4,5]

[1]Deutsche Bundesbank, Frankfurt, Germany

[2]University of Kiel, Kiel, Germany

[3]Kiel Institute for the World Economy, Kiel, Germany

[4]ifo Institute, Munich, Germany

[5]IMFS Frankfurt, Frankfurt, Germany

Correspondence
Maik H. Wolters, University of Kiel, Kiel, Kiel Institute for the World Economy, Kiel, ifo Institute, Munich, and IMFS Frankfurt, Frankfurt, Germany.
Email: wolters@economics.uni-kiel.de

Summary

Output gap revisions can be large even after many years. Real-time reliability tests might therefore be sensitive to the choice of the final output gap vintage that the real-time estimates are compared to. This is the case for the Federal Reserve's output gap. When accounting for revisions in response to the global financial crisis in the final output gap, the improvement in real-time reliability since the mid-1990s is much smaller than found by Edge and Rudd (*Review of Economics and Statistics, 2016, 98(4), 785–791*). The negative bias of real-time estimates from the 1980s has disappeared, but the size of revisions continues to be as large as the output gap itself. We systematically analyse how the real-time reliability assessment is affected through varying the final output gap vintage. We find that the largest changes are caused by output gap revisions after recessions. Economists revise their models in response to such events, leading to economically important revisions for not only the most recent years but also reaching back up to two decades. This might improve the understanding of past business cycle dynamics but decreases the reliability of real-time output gaps ex post.

KEYWORDS

business cycles, output gap, potential output, real-time data, revisions

1 | INTRODUCTION

Since the seminal paper by Orphanides and van Norden (2002), the real-time reliability of output gap estimates has been tested for many countries (see, e.g., Cayen & van Norden, 2005; Ince & Papell, 2013; Kangur et al., 2019; Marcellino & Musso, 2011). To analyse the revision properties, real-time estimates are compared to a later published output gap data vintage that is based on additional data and accounts for data revisions that have occurred in the meantime. These later output gap estimates are treated as the final revised ones.

This approach is not unproblematic since the output gap is a latent variable whose true value will never be known. Output gap revisions can be large even after several years, so that later output gap estimates can at best be a proxy for a final revised output gap. Orphanides and van Norden highlighted already in 2002 that this may be a problem: "[···] recognizing, of course, that 'final' is very much an ephemeral concept in the measurement of output" (Orphanides & van Norden, 2002, p. 571). Nevertheless, this issue has been neglected in the literature.

Based on the example of the Federal Reserve's output gap, we show that conclusions regarding the real-time reliability are indeed sensitive to the choice of the final output gap data vintage. In particular, we show that the improvement in real-time reliability of the Federal Reserve's output gap since the mid-1990s found by Edge and Rudd (2016) turns out to be substantially smaller when the currently latest available Fed output gap vintage is used as the final revised output gap.[1]

2 | SENSITIVITY OF REAL-TIME PROPERTIES OF THE FEDERAL RESERVE'S OUTPUT GAP

We use the output gap estimates that are used by Federal Reserve staff for preparing the Greenbook/Tealbook prior to the Federal Open Market Committee (FOMC) meetings (Van Cleve et al., 2019). Typically, they are made publicly available with a 5-year lag. Following the literature, we define the real-time output gap estimate for a given quarter t as the estimate from the first Greenbook of quarter $t+1$.

2.1 | Revisiting the real-time properties of the Federal Reserve's output gap

We start by focusing on real-time output gap estimates for the periods 1980Q1–1992Q4 and 1994Q1–2006Q4, which is the baseline comparison in Edge and Rudd (2016), to study changes in the real-time reliability of the Federal Reserve's output gap. They use output gap estimates that became available 2 years after the end of each sample as the final output gap estimates, that is, the ones from the November 1994 and October 2008 Greenbooks. In a first step, we compare these results to using the currently latest available output gap vintage from the December 2016 Tealbook as the final output gap estimate for both samples. Revisions are defined as the difference between the final and the real-time output gap.

Table 1 shows descriptive statistics of the output gap revisions and the final output gap estimates, alongside two noise-to-signal ratios (NSR). The NSRs are computed either as the ratio of the output gap revisions' standard deviation (SD) or root-mean-square error (RMSE), respectively, to the standard deviation of the final output gap estimate.[2] Following Edge and Rudd (2016), we assess whether changes in these NSRs are statistically significant as follows: Using a circular moving block bootstrap with a block length of 10 quarters and 5000 replications, we compute empirical distributions of the NSRs for the first sample.[3] The NSRs of the second sample are then compared to these and significance labels indicate whether they fall into the lower 1%, 5%, or 10% tail of the NSR distribution of the first sample.[4]

The upper part of the table replicates the results from Edge and Rudd (2016), showing that both NSRs decrease significantly from the first to the second sample. Based on this, Edge and Rudd conclude that the real-time reliability of the Fed's output gap has increased since the mid-1990s. The lower part shows how the results change when the latest available data vintage is used as the final estimate. Compared to before, the NSR(SD) does not decrease, but increases, though not statistically significantly. The NSR(RSME) decreases, but the decrease remains statistically insignificant. The NSR(SD) measures revision variations around the mean revision, while the NSR(RMSE) does so around the predicted revision mean of zero. The real-time output gap estimates are on average much more negative than the final estimates for the first sample, leading to a large NSR(RMSE). In the second sample, the means of the real-time and the final estimates are much closer to each other, so that the NSR(RMSE) is lower compared to the first sample. Hence, the bias of the real-time estimates from the first sample has disappeared in the second. Otherwise, the real-time reliability has not significantly changed from the first to the second sample as shown by the similarity of the NSR(SD) in the first and the second sample.[5]

[1] In addition to analysing the sensitivity of output gap real-time reliability tests with respect to the final output gap vintage choice, we have also replicated the paper by Edge and Rudd (2016) in a narrow sense as documented in Appendix S1.

[2] $NSR(SD) = \frac{\sqrt{1/(T-1)\sum_{t=1}^{T}(revision_t - mean_{revision})^2}}{\sqrt{1/(T-1)\sum_{t=1}^{T}(finalgap_t - mean_{finalgap})^2}}$ and $NSR(RMSE) = \frac{\sqrt{1/(T-1)\sum_{t=1}^{T}(revision_t - 0)^2}}{\sqrt{1/(T-1)\sum_{t=1}^{T}(finalgap_t - mean_{finalgap})^2}}$.

[3] We change the block length in the circular moving block bootstrap procedure to 10 instead of 4 or 5 as in Edge and Rudd (2016) based on insights from sample autocorrelation functions. For details, please refer to Appendix F in Appendix S1.

[4] For further details on how we draw inference, please refer also to Appendix B in Appendix S1.

[5] Output gap projections might be more relevant for forward-looking monetary policy than the $t-1$ estimates (see, e.g., Cayen & van Norden, 2005). Given that publicly available data for Greenbook/Tealbook output gap projections starts only in 1996, we cannot repeat the analysis of the real-time properties for projections of the output gap. Comparing NSRs for the subsample 1996Q1 to 2006Q4 shows only small increases in NSRs for output gap projections for $t+2$ and $t+4$ compared to the $t-1$ estimates, so that the real-time properties of output gap projections might not differ much from those of the $t-1$ estimates.

TABLE 1 Changes in real-time reliability for different final output gap vintages. [Correction added on 7 September 2023, after first online publication: The mean value of '94Q1−06Q4' for the November 1994/October 2008 'Output gap revisions' was incorrectly published as '−20.32' due to production error. This has been corrected to '−0.32' in this version.]

	Mean	SD	RMSE	NSR(SD)	NSR(RMSE)
Final output gap vintages: November 1994/October 2008					
Output gap revisions					
80Q1–92Q4	2.28	1.63	2.79	0.66	1.13
94Q1–06Q4	−0.32	0.74	0.80	0.49*	0.53***
Final gap estimates					
80Q1–92Q4	−1.64	2.47			
94Q1–06Q4	−0.11	1.52			
Final output gap vintage: December 2016					
Output gap revisions					
80Q1–92Q4	2.20	1.88	2.88	0.79	1.21
94Q1–06Q4	0.05	1.33	1.32	0.93	0.92
Final gap estimates					
80Q1–92Q4	−1.72	2.38			
94Q1–06Q4	0.26	1.44			

Note: The NSR distribution is based on a circular moving block bootstrap procedure.
Abbreviations: NSR, noise-to-signal ratio (based on SD or RMSE); RMSE, root-mean-square error; SD, standard deviation.
*The NSR of the second sample falls into the lower 10% tail of the NSR distribution in the first sample.
**The NSR of the second sample falls into the lower 5% tail of the NSR distribution in the first sample.
***The NSR of the second sample falls into the lower 1% tail of the NSR distribution in the first sample.

Orphanides and van Norden (2002) argue that NSRs capture the effects of persistent upward or downward revisions. They find that the NSR(RMSE) is larger than one for six out of eight output gap estimation methods and also for the Federal Reserve's output gap for a sample from 1966Q1 to 1997Q4, though they do not test for statistical significance. To assess whether the NRS are larger than one, we use a one-sided test based on the bootstrap distribution of the NSRs. Based on the final data vintage used in Edge and Rudd (2016), we find no significant difference from one in the first sample, but an NSR(RMSE) that is significantly smaller than one at the 5% level in the second sample. If instead the December 2016 output gap series is used as the final estimate, there is no significant difference from one in both samples.[6]

Figure 1 illustrates the underlying reasons for these results. The top panel shows the real-time output gap alongside the two alternative final output gap measures and the lower panel the two respective revision series. It becomes apparent that the real-time output gap in the first sample was highly negative throughout almost the whole 1980s, while both final output gap estimates are closer to zero on average, so that both revision series have a positive mean, correcting the negative bias of the real-time estimates. For the second sample, no such real-time bias exists, so that the NSR(SD) and NSR(RMSE) are almost equal. Independently of which final output gap measure is used, the real-time reliability has, thus, improved in the sense that the negative bias of real-time output gap estimates has disappeared since the mid-1990s. This is the reason why the NSR(RMSE) decreases for both final output gap measures.

The divergent results between the two NSRs can be explained by output gap revisions for the second sample. When the output gap from the October 2008 Greenbook is used as the final output gap, the real-time and the final output gap estimates are very similar, yielding both NSRs to decrease compared to the first sample and becoming significantly smaller than one. When the output gap from the December 2016 Tealbook is used instead, later upward revisions are included that increase the difference between the real-time and the final output gap for the period 2000–2006. Hence, the point estimate of the NSR(SD) increases, the decrease in the NSR(RMSE) becomes smaller, and the NSR(RMSE) remains insignificantly different from one.

2.2 | Which output gap revisions affect the real-time reliability?

So far, we have analysed two specific final vintage choices. Next, we systematically analyse how variations of the final vintage affect the real-time reliability analysis. Figure 2 shows how the NSR(SD) and the NSR(RMSE) vary for different final vintage choices for the two samples. Starting points at the left of each panel are given by the NSRs obtained by Edge

[6]This does not mean that the Federal Reserve's output gap does not contain useful information. Berge (2021) extracts the common component of output gap estimates from different Tealbook vintages for each point in time and shows that it contains highly valuable macroeconomic information. It is difficult to use this information in real time, though.

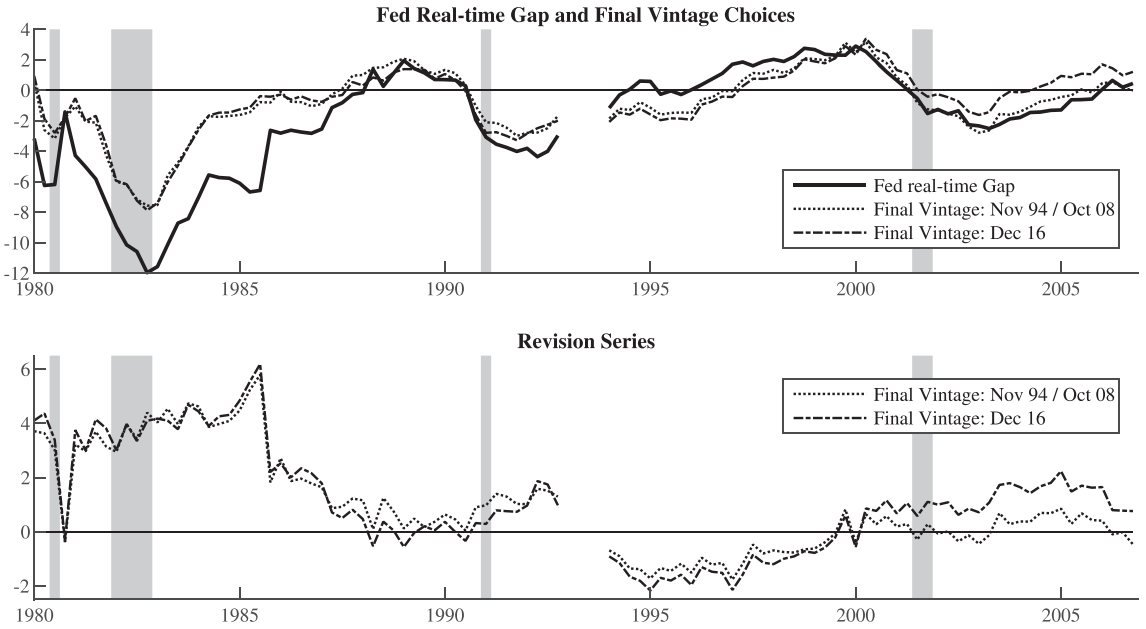

FIGURE 1 Fed real-time gap, final vintage choices, and revision series.

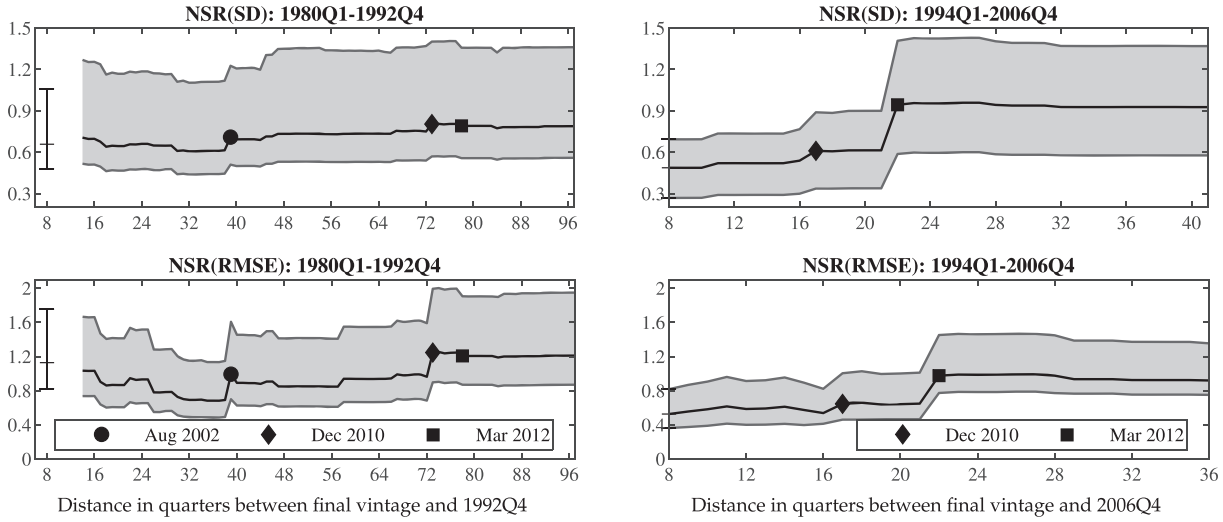

FIGURE 2 Noise-to-signal ratios for different final vintages. *Note*: The first row shows the evolution of the NSR(SD) and the second the evolution of the NSR(RMSE) when shifting the vintage used for measuring the final output gap quarter-by-quarter. The grey area represents 90% confidence bands, based on the 5th and 95th percentile of the NSR bootstrap distributions. The markers indicate the three largest revisions.

and Rudd (2016). The publication dates of the output gap vintages used as the final output gaps are eight quarters after the last observation of the two respective real-time samples as denoted on the horizontal axis. The last NSRs in all graphs are the ones obtained from using the December 2016 Tealbook as final revised output gap, so that this amounts to the maximum possible distance between the publication of the real-time estimates and the final ones. In between these two choices, we show the effect of all other possible final output gap choices on the NSRs by moving quarter-by-quarter from one possible final output gap vintage to the next.

Fed output gap vintages that include long enough time series are publicly available since the June 1996 Greenbook, with an additional vintage from the November 1994 Greenbook being available from the replication material of Edge and Rudd (2016) who obtained it from Athanasios Orphanides. Hence, in the graph on the left, we include the NSRs from Edge and Rudd's first sample that are based on the November 1994 vintage, leave a gap for the missing vintages, and then

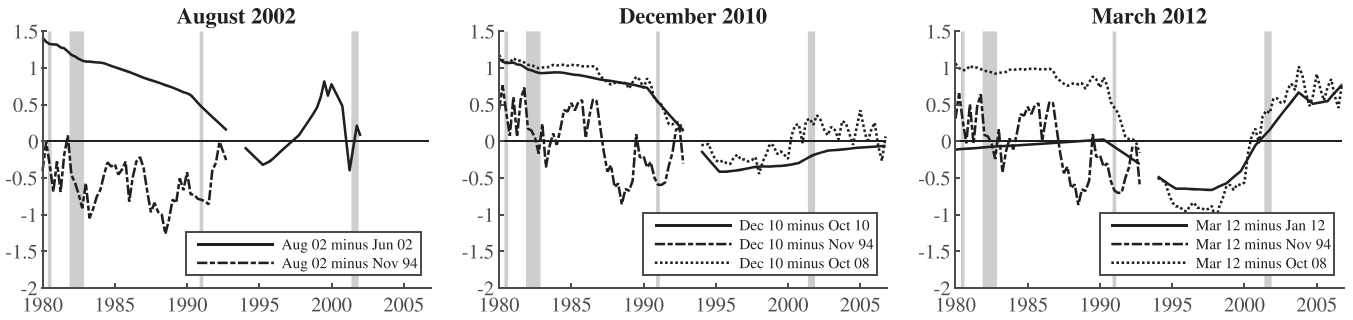

FIGURE 3 Three largest output gap revisions. *Note*: The panels show output gap revisions to the vintage from the previous Greenbook in percentage points (solid) and to the final vintage choice from Edge and Rudd (November 1994: Dash-dotted; October 2008: Dotted) to the output gap vintages from August 2002, December 2010, and March 2012. Grey bars indicate NBER recessions.

show how the NSRs change when shifting the final vintage quarter-by-quarter starting with the June 1996 Greenbook. In the graph on the right, we start with the NSRs from Edge and Rudd's second sample based on the December 2008 vintage and then shift the final vintage quarter-by-quarter.

In both samples, the NSRs show substantial variations when varying the final vintage, so that robustness checks seem to be generally necessary for real-time reliability analyses.[7] While there is a slight overall upward tendency when later vintages are used as final vintage, changes in results are mainly driven by three revisions. In the first sample, the revisions from the August 2002 Greenbook and the December 2010 Tealbook have the largest impact that is best visible for the NSR(RMSE). Further, a number of smaller revisions between 1994 and 2000 have in sum a sizable effect on the NSR(RMSE). The change in the NSRs via varying the final vintage is even larger in the second sample. The revision from the March 2012 Tealbook increases both NSRs strongly. While the changes in the NSRs in the first sample do not significantly change the results reported in Edge and Rudd (2016), the March 2012 revision leads to a significant increase in both NSRs in the second sample.

Figure 3 shows these three largest output gap revisions. Each plot shows the output gap revision to the vintage from the previous Greenbook in percentage points as well as the revisions to the final estimates used by Edge and Rudd. The left panel shows that following the 2001 recession, the Fed revised potential output downwards leading to a large upward revision of the output gap from the June to the August 2002 Greenbook. The output gap revisions are largest for the pre-recession period and for observations in the early 1980s. The upward revision of the 1980s output gap corrects the large negative bias of the real-time estimates in the first sample and therefore increases the NSR(RMSE) more than the NSR(SD). Edge and Rudd's results for the first sample are not much affected though because this upward revision of the output gap offsets a series of previous downward revisions, so that the difference to the November 1994 vintage is not as large.

Following the global financial crisis, two large output gap revisions occurred. The first from the December 2010 Tealbook, shown in the middle panel, leads to a large upward revision of the output gap for the first sample. Similar to the August 2002 revision, this revision does not affect the results by Edge and Rudd as it offsets downward revisions in the Greenbooks between 1994 and 2000. The revisions for the second sample are negligible in comparison to those for the first sample, so that the NSRs are virtually unchanged to using the October 2008 vintage used by Edge and Rudd. The second revision after the global financial crisis occurred in the March 2012 Tealbook and is shown in the right panel. The Fed revised potential output downwards, leading to an upward revision of the output gap not only around the Great Recession and the pre-recession period, but going back to 2001. For the period 1994–2000, there is an upward revision of potential output leading to a downward revision of the output gap. Together, these revisions lead to large increases in the NSRs in the second sample and the divergence of the results documented in Table 1.[8] The first sample is hardly affected by the March 2012 revision.

[7]The changes of the NSRs are driven by changes in the SD or RMSE of the revisions, respectively, while the SD of the final output gap, that is, the denominator of the NSRs, remains almost perfectly constant for different final output gap vintages.

[8]The Tealbook remains confidential for 5 years, so that the large impact of the March 2012 revision could have only been known at the time for those having access to the Tealbook output gap. For outsiders, the impact could not have been known before 2017, that is, 1 year after the publication of Edge and Rudd (2016). Surprisingly, so far, neither the original authors nor any other researchers have updated the results by Edge and Rudd (2016) to check whether the results continue to hold.

2.3 | Which vintage to use as the final output gap estimate?

The various revisions lead to the question which vintage to choose as the final one. On the one hand, the most recently available vintage should contain the most complete information regarding past business cycles for the following three reasons: First, the increasing number of available observations, second, the incorporation of data revisions, and, third, the revision of modelling approaches following events like recessions, crises, or periods of strong productivity growth like in the mid-to-late 1990s. On the other hand, there might be output gap revisions for consistency reasons that do not necessarily improve the measurement of past business cycles. For example, if a trend is adjusted based on observations for recent years without checking whether this necessitates the modelling of a trend break, the application of the same trend to observations in the more distant past might distort their business cycle measurement.

It is a priori not clear which argument dominates, so that robustness checks and a case-specific justification of the final vintage choice are advisable. In the following, we undertake a cautious attempt of such a justification. All three major output gap revisions shown in Figure 3 occur after recessions. Revisions around recessions are likely based on new information leading economists to update their models.[9] These revisions are also explained in the Tealbook, so that it seems plausible to take them into account when evaluating real-time output gap estimates. Revisions of the distant past are, on the other hand, unlikely to be directly related to recent developments and there is no explanation regarding these earlier revisions in the Tealbook.[10]

Based on these considerations, we propose to use the corresponding observation from the first output gap vintage after the subsequent NBER-defined recession that entails a substantial revision as the final output gap estimate for a specific real-time output gap observation. In this way, new information based on the recession that led to the revision is included, but later revisions that do not necessarily improve past output gap estimates are omitted. This approach can only be implemented approximately, because output gap vintages before 1994 are not available. We, thus, compare the 1980s real-time output gap to the 1994 vintage as the final revised estimate. For the real-time output gaps after the early 1990s recession until the end of the 2001 recession, we then use the output gap from the August 2002 Greenbook. Lastly, we use the output gap from the March 2012 Tealbook to incorporate the first large revision after the Great Recession for the real-time observations after the 2001 recession until the end of the sample in 2006.

Using this approach, we find no significant change in the NSR(SD) from the first to the second sample. The decrease in the NSR(RMSE) is less significant than reported in Edge and Rudd (2016). Further, the null hypothesis that the NSR(RMSE) is equal to one cannot be rejected in both samples. The results are, thus, similar to using the December 2016 Tealbook as the final revised vintage. This is not surprising since the March 2012 revision that causes the divergence in the results documented in Table 1 is included.[11]

3 | SENSITIVITY OF REAL-TIME PROPERTIES OF STATISTICAL OUTPUT GAP ESTIMATES

Are statistical gap estimates similarly prone to the final vintage choice? To analyse this, we use real-time GDP data vintages to compute real-time estimates for 11 popular univariate output gap estimation methods including deterministic detrending methods, standard univariate unobserved component models, Beveridge-Nelson-based approaches as well as bandpass and regression-based filtering techniques.[12]

[9]For many statistical models, upward revisions as in 2002, 2010, and 2012 can be expected as relatively volatile negative shocks of a preceding recession are propagated by the linear weights of a Kalman smoother.

[10]Historical Greenbooks are available on https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm. The August 2002 downward revision of potential output for the late 1990s and early 2000s is justified in the Greenbook with the downward revision of productivity growth in the national accounts for this period. By contrast, the large revisions of the output gap during the 1980s are not mentioned. The December 2010 Tealbook does not mention any revisions at all. In the March 2012 Tealbook, the downward revision of potential GDP for the years 1996 to 2010 is explained. A lower level of potential output is more in line with the Fed's estimate of labour market slack at the end of 2011. The reasons for the earlier downward revision from 1994 to 2001 remain unclear, though.

[11]Another final vintage choice could be based on testing which output gap vintage provides the best inflation forecast, that is, contains the information that is desirable for monetary policy. Differences in inflation forecasting accuracy are tiny when using different output gap vintages, though. Hence, this approach is not applicable in practice.

[12]Specifically, we include linear detrending, linear detrending with trend breaks (end of 1973 and beginning of 1997 assumed to be known with a 3-year delay), quadratic detrending, the univariate unobserved components model by Watson (1986) with a random walk with constant drift and an AR(2) cycle, the extension by Harvey (1985) and Clark (1987) allowing for time-variation of the drift term, the Hodrick-Prescott filter, the Baxter-King bandpass

TABLE 2 Noise-to-signal ratios from statistical output gap estimates.

	Linear trend	Broken trend	Quadratic trend	Watson	Harvey-Clark	Hodrick-Prescott	Baxter-King	Beveridge-Nelson	Modified Bev.-Nel.	Hamilton	Modified Hamilton	Memo: Fed
NSR(SD)												
Final output gap vintages: November 1994/October 2008												
80Q1–92Q4	0.64	0.56	0.47	0.36	0.66	1.08	0.76	0.55	0.24	0.17	0.17	0.66
94Q1–06Q4	1.25**	1.16***	0.61**	0.47**	0.67	1.11	0.57	0.72	0.54***	0.45***	0.45***	0.49*
Final output gap vintage: December 2016												
80Q1–92Q4	0.37	0.55	0.50	0.43	0.67	1.07	0.68	0.52	0.30	0.22	0.20	0.79
94Q1–06Q4	0.79***	1.04***	0.85***	0.69***	0.85**	1.04	0.72	0.72*	0.52**	0.47***	0.46***	0.93
NSR(RMSE)												
Final output gap vintages: November 1994/October 2008												
80Q1–92Q4	1.48	0.60	0.57	0.74	0.65	1.08	0.83	0.55	0.25	0.18	0.18	1.13
94Q1–06Q4	1.85	1.15***	1.47***	0.98*	0.69	1.16	0.57	0.72	0.56**	0.55***	0.56***	0.53***
Final output gap vintage: December 2016												
80Q1–92Q4	2.96	0.59	1.06	1.84	0.67	1.08	0.78	0.55	0.43	0.24	0.24	1.21
94Q1–06Q4	4.80*	1.04***	0.84**	3.08*	0.98**	1.07	0.73	0.71	0.56	0.51***	0.54**	0.92

Note: See Table 1.

*The NSR of the second sample falls into the lower 10% tail of the NSR distribution in the first sample.

**The NSR of the second sample falls into the lower 5% tail of the NSR distribution in the first sample.

***The NSR of the second sample falls into the lower 1% tail of the NSR distribution in the first sample.

For all methods, we start the estimation in 1954Q1. We compute NSRs for the case when the final output gap estimates are based on GDP data available in November 1994 and October 2008, respectively, and for the case when they are based on data available in December 2016. Table 2 shows the NSR(SD) (upper part) and the NSR(RMSE) (lower part) for the two samples. As before, significance labels indicate whether the NSRs of the second sample are significantly different from the ones in the first sample.[13]

For the linear and the quadratic deterministic detrending, the NSRs change when using a later output gap vintage as the final estimate. In light of the 1970s growth slowdown and the 1990s growth acceleration that affect the deterministic trend, this is not surprising. The bias of these real-time estimates can be avoided by allowing for breaks in the trend as shown in the column "Broken Trend." Also the Watson model, in which the drift term is assumed to be constant, is sensitive to the final vintage choice. For the other methods, we find little or no sensitivity with respect to the final vintage choice. That does not mean that the methods are not prone to an end-point problem, but the revisions in the first two years after the sample end are already sufficient to get reliable output gap estimates that are not further revised when using later data vintages. These results confirm that the output gap revisions of the Fed in August 2002, December 2010, and March 2012 are based on judgement of the Federal Reserve staff that goes beyond applying standard statistical trend-cycle decompositions.[14]

The results confirm Edge and Rudd's (2016) finding that there is no reduction in the NSRs from the first sample to the second sample for statistical output gap estimates. A comparison to the NSRs of the Fed's output gap (last column of Table 2) reveals, however, that part of the reduction found by Edge and Rudd stems from the particularly high unreliability in the first sample, when its output gap estimate was characterized by a large negative real-time bias. Since both NSRs are very similar for most statistical output gap estimates, they were, in contrast, not characterized by such large real-time biases. Further, when the December 2016 vintage is used as the final revised output gap, the Fed's output gap NSRs are even at the upper end of the range implied by statistical output gap estimates in the second sample. Some statistical output gap estimates yield NSRs that are significantly smaller than one in both samples for both final vintage choices (Harvey-Clark, Baxter-King, the two Beveridge-Nelson decompositions, and the two Hamilton-based filters). The NSR of the Fed's output gap is only significantly smaller than one in the second sample when using Edge and Rudd's (2016) final vintage choice. On the other hand, most statistical output gap measures show an increase in the NSRs from the first to the second sample. Hence, while the significant decrease of the NSR of the Fed's output gap disappears when the December 2016 final vintage is used, there is still the possibility that there has been an improvement in the Fed's ability to produce reliable output gaps, which may have been offset by a more difficult environment.[15]

4 | ROBUSTNESS AND UPDATE

Next, we check the robustness of the results for alternative samples and how the results are affected when we extend the analysis and include the latest currently publicly available output gap vintages.[16]

filter, a Beveridge-Nelson decomposition based on an ARIMA(1, 1, 0) process for log real GDP, the modified Beveridge-Nelson decomposition by Kamber et al. (2018) producing a more persistent cyclical component with higher amplitude, the one-sided regression-based filter by Hamilton (2018), and the modified version by Quast and Wolters (2022) that covers typical business cycle frequencies more evenly and yields a smooth trend estimate.

[13]Results for the Watson and the Harvey-Clark model reported in the upper part of the table differ somewhat from Edge and Rudd (2016), because we use the MCMC approach of (Grant & Chan, 2017a, 2017b) that avoids implausibly large variations in the NSRs when varying the final output gap vintage.

[14]We also shifted the final vintage for the statistical output gap estimates quarter-by-quarter as shown in Figure 2 for the Fed's output gap and did not observe any sizable changes in the NSRs except for deterministic detrending and the Watson model with constant drift term.

[15]Table 2 also shows large differences in NSRs between methods. It is important to note that the NSRs are only informative with respect to revision size, but do not contain any information on which output gap is preferable based on its economically meaningfulness (see, e.g.,Barbarino et al., 2020; Quast & Wolters, 2022, for such analyses). Further, we restrict the analysis to univariate methods, while the Fed's analysis starts from the labour market, so that labour market indicators and GDP are accounted for (see, e.g., Fleischman & Roberts, 2011). Additionally, recent contributions often focus on multivariate methods (see, e.g., Barigozzi & Luciani, 2021; Hasenzagl et al., 2022; Jarocinski & Lenza, 2018; Morley & Wong, 2020). Work by Federal Reserve Board economists (Barbarino et al., 2020) emphasizes the importance of including the unemployment rate in output gap models to increase the real-time reliability.

[16]Tables with detailed results for this section are available in Appendix S1.

4.1 | Robustness

We repeat the analysis discussed in Section 2 for the samples 1966Q1–1997Q4 and 1998Q1–2006Q4 as in Edge and Rudd (2016). For this specification, we find a significant increase in the real-time reliability of the Fed's output gap for both final vintage choices, but the results are still sensitive to the final vintage choice. The point estimates of the NSRs decrease less when the December 2016 vintage is used as the final output gap. Further, while Edge and Rudd's (2016) final vintage choice yields an NSR(RMSE) significantly smaller than one in the second sample, this is not the case when using the December 2016 vintage as the final output gap. While there is some evidence that the real-time reliability has improved compared to this longer first sample, this is not so surprising because this sample choice makes it more likely to find improvements in the Fed's output gap estimation as argued by Edge and Rudd (2016). First, the shorter second sample leaves less time for revisions to occur compared to the longer first sample. Second, the longer first sample includes the 1970s growth slowdown, while the mid-1990s growth acceleration is partially excluded from the second sample. Third, in the 1960s and 1970s, the Fed did not estimate potential output itself, but relied on the estimates from the Council of Economic Advisers.[17]

4.2 | Update

With additional output gap vintages having become publicly available since Edge and Rudd's (2016) work was published, we update the results to analyse how the real-time reliability of the Fed's output gap has developed during the global financial crisis and zero lower bound period. We use real-time estimates for the output gap from 2007Q1 to 2014Q4 and use the output gap from the December 2016 Tealbook as the final revised estimate.[18] The point estimates of the NSRs are small (NSR(SD): 0.38, NSR(RMSE): 0.57), but estimation uncertainty remains relatively high. Only the NSR(SD) is significantly smaller than one at the 5% level. Compared to the 1980–1994 sample, there is a significant decrease in the NSRs regardless of the final vintage definition on which the NSRs' computations for the first sample are based on. Compared to the 1996–2006 sample, there is no significant decrease of the NSR(RMSE) when the 2008 final vintage is used for the previous sample, but a highly significant decrease, when the 2016 final vintage is used instead. These results are in line with Barbarino et al. (2020) who find an improvement in the real-time reliability of the Fed's output gap for a sample that includes the Great Recession. Since we find that the NSRs of statistical output gap estimates decrease as well when including the Great Recession, the economic environment might have changed in a way that output gap revisions have generally become smaller.

5 | CONCLUSION

We have shown that real-time reliability assessments of output gaps can be sensitive to the choice of the final output gap vintage. In particular, the real-time reliability improvements in the Federal Reserve's output gap estimates since the mid-1990s found in previous work are much lower when revisions in response to the global financial crisis are taken into account. After the Great Recession, the Federal Reserve staff revised its assessment of potential output, affecting output gap estimates since 1990. Accounting for this revision in the final output gap estimate, we find that revisions of the Federal Reserve's output gap continue to be of the same order of magnitude as the output gap itself. We do not observe such a large revision for statistical output gap estimates, so that the revision is the result of economic considerations that include a substantial amount of economic expertise rather than being based only on statistical models. Hence, these revisions might indeed reflect an improvement of the Federal Reserve's understanding of past business cycles. We show that the choice of the final output gap vintage is not only crucial with respect to the output gap revisions after the Great Recession. We detect several other output gap revisions that occur after recessions and affect the real-time reliability of the Federal Reserve's output gap. Hence, when analysing the real-time reliability of output gaps, it is important to check

[17] We also check whether this alternative sample choice affects the results discussed in Section 3. As in the baseline specification, the real-time reliability of the statistical output gap estimates is insensitive with respect to the final vintage choice, except for deterministic detrending and the Watson model.

[18] Results for the statistical output gap estimates considered in Section 3 are provided in Appendix S1.

how the results depend on the choice of the final output gap vintage. Even if one finds small revisions for a given sample, conclusions that an output gap has a high real-time reliability might be premature, because limiting the period over which one observes revisions may bias estimates of reliability. Results may change should the output gap be revised further in future.

ACKNOWLEDGEMENTS

We thank Heather Anderson, two anonymous referees as well as Joshua Chan, Jonas Dovern, Philipp Hauber, Nils Jannsen, Vivien Lewis, Elmar Mertens, Jeremy Rudd, Yves Schüler, and Simon van Norden for valuable comments. The views in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Deutsche Bundesbank or the Eurosystem.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.15456/jae.2023060.0836027355.

REFERENCES

Barbarino, A., Berge, T. J., Chen, H., & Stella, A. (2020). Which output gap estimates are stable in real time and why? Finance and Economics Discussion Series 2020-10.

Barigozzi, M., & Luciani, M. (2021). Measuring the output gap using large datasets. *The Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest_a_01119

Berge, T. J. (2021). Time-varying uncertainty of the Federal Reserve's output gap estimate. *The Review of Economics and Statistics*, 1–38. https://doi.org/10.17016/FEDS.2020.012r1

Cayen, J.-P., & van Norden, S. (2005). The reliability of Canadian output-gap estimates. *North American Journal of Economics and Finance*, *16*(3), 373–393.

Clark, P. K. (1987). The cyclical component of US economic activity. *The Quarterly Journal of Economics*, *102*(4), 797–814.

Edge, R. M., & Rudd, J. B. (2016). Real-time properties of the Federal Reserve's output gap. *Review of Economics and Statistics*, *98*(4), 785–791.

Fleischman, C. A., & Roberts, J. M. (2011). From many series, one cycle: Improved estimates of the business cycle from a multivariate unobserved components model. Finance and Economics Discussion Series 2011-046, Board of Governors of the Federal Reserve System.

Grant, A. L., & Chan, J. C. (2017a). A Bayesian model comparison for trend-cycle decompositions of output. *Journal of Money, Credit and Banking*, *49*(2-3), 525–552.

Grant, A. L., & Chan, J. C. (2017b). Reconciling output gaps: Unobserved components model and Hodrick-Prescott filter. *Journal of Economic Dynamics and Control*, *75*, 114–121.

Hamilton, J. D. (2018). Why you should never use the Hodrick-Prescott filter. *Review of Economics and Statistics*, *100*(5), 831–843.

Harvey, A. C. (1985). Trends and cycles in macroeconomics time series. *Journal of Business and Economic Statistics*, *3*(3), 216–227.

Hasenzagl, T., Pellegrino, F., Reichlin, L., & Ricco, G. (2022). A model of the Fed's view on inflation. *The Review of Economics and Statistics*, *104*(4), 1–19.

Ince, O., & Papell, D. H. (2013). The (un)reliability of real-time output-gap estimates with revised data. *Economic Modelling*, *33*, 713–721.

Jarocinski, M., & Lenza, M. (2018). An inflation-predicting measure of the output gap in the euro area. *Journal of Money Credit and Banking*, *50*(6), 1189–1224.

Kamber, G., Morley, J., & Wong, B. (2018). Intuitive and reliable estimates of the output gap from a Beveridge-Nelson filter. *Review of Economics and Statistics*, *100*(3), 550–566.

Kangur, M. A., Kirabaeva, K., Natal, J.-M., & Voigts, S. (2019). How informative are real time output gap estimates in Europe? (*IMF Working Paper WP/19/200*).

Marcellino, M., & Musso, A. (2011). The reliability of real-time estimates of the Euro Area output gap. *Economic Modelling*, *28*(4), 1842–1856.

Morley, J., & Wong, B. (2020). Estimating and accounting for the output gap with large Bayesian vector autoregressions. *Journal of Applied Econometrics*, *35*(1), 1–18.

Orphanides, A., & van Norden, S. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, *84*(4), 569–583.

Quast, J., & Wolters, M. H. (2022). Reliable real-time output gap estimates based on a modified Hamilton filter. *Journal of Business and Economic Statistics*, *40*(1), 152–168.

Van Cleve, L., Laforte, J.-P., & Stella, A. (2019). Real-time historical estimates of the output gap. FEDS Notes No. 2019-10-15. https://doi.org/10.17016/2380-7172.2460

Watson, M. W. (1986). Univariate detrending methods with stochastic trends. *Journal of Monetary Economics*, *18*(1), 49–75.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Quast, J., & Wolters, M. H. (2023). The Federal Reserve's output gap: The unreliability of real-time reliability tests. *Journal of Applied Econometrics*, *38*(7), 1101–1111. https://doi.org/10.1002/jae.3003