Measurement Error in Macroeconomic Data and Economics Research: Data Revisions, Gross Domestic Product, and Gross Domestic Income

Andrew C. Chang* and Phillip Li†

October 10, 2017

Abstract

We use a preanalysis plan to analyze the effect of measurement error on economics research using the fact that the Bureau of Economic Analysis (BEA) both revises its gross domestic product (GDP) data and also publishes a second, theoretically identical estimate of US output that only differs from GDP due to measurement error: gross domestic income (GDI). Using a sample of 23 models published in top economics journals, we find that reestimating models using revised GDP always gives the same qualitative result as the original publication. Estimating models using GDI instead of GDP gives a different qualitative result for 3 of 23 models (13%).

JEL Codes: C80; C82; E01

Keywords: Data Revisions; File-Drawer Problem; Gross Domestic Product; GDP; Gross Domestic Income; GDI; Latent Output; Measurement Error; National Income and Product Accounts; NIPA; Real-Time Data; Preanalysis Plan; Publication Bias

---

*Chang: Senior Economist, Division of Research and Statistics, Board of Governors of the Federal Reserve System. 20th St. NW and Constitution Ave., Washington DC 20551 USA. +1 (657) 464-3286. a.christopher.chang@gmail.com. https://sites.google.com/site/andrewchristopherchang/

†Li: Researcher, Office of Financial Research, U.S. Department of the Treasury. 717 14th St. NW, Washington DC 20005 USA. phillip.li@ofr.treasury.gov

‡The views and opinions expressed here are those of the authors and are not necessarily those of the Board of Governors of the Federal Reserve System, the Department of the Treasury, or the Office of Financial Research. This paper was started before Li joined the Office of Financial Research. An earlier version of this paper is in the Board's Finance and Economics Discussion Series as 2015-102, Chang and Li (2015b). We thank anonymous referees, Stephanie Aaronson, David F. Hendry, Kirstin Hubrich, Evan F. Koenig, Jeremy J. Nalewaik, Christopher J. Nekarda, Marie-Noëlle Robitaille, Bo Sun, Kurt von Tish, and seminar participants at the 2016 APPAM annual meeting, the Board, the Canadian Economics Association 2016 annual meeting, FRB Cleveland, GWU, OCC, OFR, Penn State, SJSU, UC - Irvine, and the 2016 WEAI annual meeting for helpful comments. We thank Tyler J. Hanson, Erik Larsson, Kim T. Mai, Anthony Marcozzi, Shawn M. Martin, Tyler Radler, Adam Scherling, and John Stromme for research assistance. We also thank Felix Galbis-Reig and Spencer C. Li for technical assistance. Any errors are ours.

1 Introduction

Low unemployment. Modest inflation. High output growth. Economists have devoted a substantial amount of effort to understanding how countries can achieve these three goals. Unfortunately, the unemployment rate, the inflation rate, and the output growth rate of an economy are all unobserved variables. Econometricians rely on estimates of these unobserved variables from national statistical agencies. For example, to estimate the unobserved output growth rate of the US economy, the Bureau of Economic Analysis (BEA) publishes US gross domestic product (GDP). But because published GDP is based on finite samples and imperfect source data, published GDP contains measurement error.

Using a preanalysis plan, our paper analyzes the potential effect that the measurement error in US GDP has on economics research. In addition to the more well-known data revision dimension, where the BEA revises previously released statistics to incorporate better methodologies or source data, our preanalysis plan exploits the fact that the BEA also publishes two theoretically identical measures of unobserved US output. First, the BEA publishes the more familiar GDP measure of unobserved output that estimates it based on total expenditures. Second, the BEA publishes the less familiar gross domestic income (GDI) measure that estimates it based on total income. As total expenditures must equal total income, GDP and GDI are theoretically identical. But because of measurement error the BEA's published GDP and GDI statistics differ.

Our analysis of the potential effect of measurement error in GDP on economics research proceeds in three steps. First, we acquire a sample of 67 published economics papers that use US GDP to estimate a key result, and we replicate 29 of these published papers (Chang and Li, 2015a, Forthcoming). Second, using the original replication data files we identify which vintage (publication date by the BEA) of GDP the published papers use in their estimation by comparing the data files to historical vintages of GDP. We identify the original data vintage for 23 papers. Third, we reestimate these 23 papers by replacing the original vintage of GDP the authors use with the original vintage of GDI, the revised current-vintage GDP, and the revised current-vintage GDI.

Comparing the key results of the 23 published papers to the results we find using the three alternative estimates of unobserved output (current-vintage GDP, original-vintage GDI, and current-vintage GDI), we find that current-vintage GDP gives the same qualitative result as the original article in all 23 cases. But when we estimate models with either original-vintage GDI or current-vintage GDI, the results we obtain exhibit greater quantitative differences from the published results compared with when we estimate the models with current-vintage GDP. For three papers, the results we obtain when using GDI (both with original-vintage GDI and current-vintage GDI) are qualitatively different than the original articles.

This paper has three main contributions to the literature on measurement error and national statistics.[1]

For our first contribution, we analyze the effect of measurement error on 23 papers sampled from 11 top economics journals, which is a larger and more comprehensive set of economics papers than the literature has used. Previous studies that look at the effect of measurement error on economics research typically select a single paper or use a selected small sample of papers to highlight their claims.[2] Our broad sample mitigates selection bias concerns.

For our second contribution, we contrast the effects of measurement error across both revisions to the same estimate of unobserved output (GDP) and also against a theoretically identical estimate of unobserved output based on completely different source data (GDI).

[1]Examples of studies that look into measurement error of national statistics include Mankiw and Shapiro (1986); Orphanides (2001); Orphanides and van Norden (2002); Koenig, Dolmas, and Piger (2003); Nalewaik (2010); Ponomareva and Katayama (2010); Wolff, Chong, and Auffhammer (2011); Feng and Hu (2013); Zucman (2013); Nalewaik (2014) and Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2016).

[2]Four examples of research that looks at whether measurement error affects a single paper or a small sample of papers include: (1) Ponomareva and Katayama (2010), who use Ramey and Ramey (1995) as an example of how revisions to the Penn World Tables may influence results, (2) Wolff, Chong, and Auffhammer (2011), who use Noorbakhsh (2006) as an example of how results change due to measurement error in the human development index released by the United Nations Development Programme, (3) Croushore and Stark (2002, 2003), which our study is closest to, who examine the effect of data revisions on the qualitative results of Hall (1978); Blanchard and Quah (1989) and Kydland and Prescott (1990), and (4) Faust, Rogers, and Wright (2003), who use many data vintages to analyze the exchange rate forecasting model of Mark (1995).

To our knowledge, our paper is the first to investigate the effects of measurement error in a national statistic where a second, theoretically identical measure of the same quantity of interest is available.

Under normal circumstances when two estimates of the same quantity of interest are available, the estimates use different data definitions. For example, the current employment statistics (CES) and the current population survey (CPS) both estimate US employment. The CES and the CPS, however, have different definitions of what it means to be employed.[3] Partly because of different data definitions, total employment in the CES differs from total employment in the CPS. The differences between the CES and CPS are not just because of measurement error. Economic models generally abstract from different data definitions. In this paper, we use the fact that GDP and GDI only differ because of measurement error as both GDP and GDI are estimates of the same quantity: unobserved output. As far as we are aware, the existing literature on the effect of measurement error of national statistics on economic research focuses only on revisions to the same statistic or survey dataset.[4] The purpose of this paper is to document what the effect on economic research would be when using different measures of latent output in estimated models. This paper does not investigate whether GDP or GDI is a better measure of latent output nor do we analyze what qualities of GDP or GDI lead models to give different results.

Our third contribution is that we use a preanalysis plan, which maximizes both the credibility and transparency of our findings, to investigate the relationship between measurement error, data revisions, and economics research. To our knowledge, our paper is also the first paper to use a preanalysis plan to investigate this relationship. While prolific in the medical

[3]The CES measure of job gains uses changes in private payroll employment, where each new job is from the establishment-side perspective. In the CES individuals with multiple jobs are counted at each establishment the individual is employed at. The CPS measures job gains by household-level employment where each new employed individual counts as a new job. Therefore, in the CPS individuals with multiple jobs count as one employed person.

[4]For example, Croushore and Stark (2003); Koenig, Dolmas, and Piger (2003); Ponomareva and Katayama (2010) and Wolff, Chong, and Auffhammer (2011). For a review of research into real-time data, see Croushore (2011).

field, preanalysis plans in economics are still relatively uncommon.[5]

2 Description of BEA's Data Release Schedule, GDP, and GDI

We provide a brief description of the BEA's data release schedule, GDP, and GDI. Interested readers can see Fixler and Grimm (2008) or Landefeld, Seskin, and Fraumeni (2008) for additional details on the BEA's construction of GDP and GDI. The appendices to Nalewaik (2010) provide information on the source data behind GDP and GDI.

The BEA publishes its first release of GDP for the previous quarter, called the advance release, about one month after the quarter ends. The BEA then publishes a once-revised release for the previous quarter, called the second release, in the next month and a twice-revised release for the previous quarter, called the third release, another month thereafter. For example, the advance release of Q4 GDP would appear in January, the second release would appear in February, and the third release would appear in March. The third release is then unrevised until the summer, when the BEA conducts an annual revision and revises its published statistics for the last three calendar years. In addition, once about every five years, the BEA conducts a comprehensive revision where all of its previously published statistics are potentially revised.[6]

Figure 1 plots the net revision to the level of nominal GDP between the September 26th, 2013 vintage of GDP and the September 26th, 2008 vintage of GDP.[7] Figure 1 shows that the BEA revised up the level of GDP from the September 26th, 2008 vintage. In addition, from

[5] The earliest paper in economics, to our knowledge, to use a preanalysis plan is Neumark (1999, 2001) who investigates the effect of minimum wages on employment. More recent examples of research that use preanalysis plans include Casey, Glennerster, and Miguel (2012) and Chang and Li (2015a, 2017, Forthcoming).

[6] The last three comprehensive revision estimates were released on December 10th, 2003; July 31, 2009; and July 31, 2013.

[7] Because of chain aggregation and differences in the GDP deflator, comparisons of real GDP across BEA vintages are not meaningful. See Whelan (2002) for a discussion. We choose an arbitrary difference of five years to illustrate the statistical discrepancy.

2008 to 2013 the BEA's revised published GDP estimates all the way back to the 1940s.[8]

Because GDP is nonstationary and economic models generally take nonstationarity into account, a more informative version of Figure 1 may be a comparison of vintages of a stationary transformation of GDP. Figure 2 plots the net revision to nominal GDP between the same two vintages shown in Figure 1, except expressed as annual percent changes to make the GDP series stationary. Figure 2 shows that the revisions to the annual percent changes of GDP tend to be larger for more recent data. Yet even for data that have already undergone a comprehensive revision, subsequent revisions can have meaningful changes on published estimates. For example, published estimates of GDP growth of the 1990s were often revised by $\pm 0.5\%$ due to revisions that occurred between 2008 and 2013, about ten years after the initial GDP estimates were published. The magnitude of the average net revision from 2008 to 2013 of GDP for 1947 to 2008 is 0.30 percentage point.[9]

GDP is an estimate of unobserved output, as defined as the total value of goods and services produced in the economy, that the BEA creates using data on expenditures. At a high level, this approach corresponds to using data on consumption, investment, government expenditures, and net exports (Landefeld, Seskin, and Fraumeni, 2008). We emphasize that the BEA's published GDP is an estimate of the total value of goods and services produced in the economy based on expenditure-side data. Published GDP is generally not the actual total value of goods and services produced in the economy, which is generally the unobserved output variable of interest to economists.

The release schedule for GDI is similar to the schedule for GDP, although the data the BEA uses to construct GDI are less timely than the data for GDP.[10] The BEA publishes

[8]The BEA revised up the level of GDP in the 2013 comprehensive revision to allow for capitalization of R&D, among other changes to the national accounts. If a researcher considers older vintages of a statistic to be more mismeasured than newer vintages, than the pre-2013 comprehensive revision vintages of GDP were all mismeasured because they did not take into account this capitalization of R&D.

[9]For GDP since 1990 the magnitude of the average net revision is 0.59 percentage point. For GDP since 2000 the magnitude of the average net revision is 0.82 percentage point. The magnitude of net revisions to GDP in NBER expansions since 1947 Q1 (0.29 percentage points) is about the same as in NBER recessions since 1947 Q1 (0.33 percentage points), although since 1990 Q1 the net revisions in recessions have been larger on average.

[10]The timeliness of the source data is one reason the BEA prefers GDP to GDI (Landefeld, Seskin, and

its first release of GDI about two months after the quarter ends, except in the case of the fourth quarter, when it publishes its first GDI statistic three months after the quarter ends. For the first through third quarters, the BEA revises its initial GDI release a month after publication. The BEA's GDI releases are subject to the same annual and comprehensive revision schedule as GDP.

GDI is an estimate of unobserved output the BEA produces using data on income that includes: compensation, rental income, profits and proprietor's income, taxes less subsidies, interest, miscellaneous payments, and depreciation (Landefeld, Seskin, and Fraumeni, 2008). Like GDP, the BEA's published GDI is an estimate of unobserved output, not the unobserved output variable of interest to economists.

In theory, GDP and GDI should be identical. Both GDP and GDI are estimates of the same quantity: unobserved output. But because the data the BEA uses to construct GDP and GDI are imperfect and largely independent, the published estimates of GDP and GDI differ from each other and contain measurement error. The BEA refers to the difference between GDP and GDI as the statistical discrepancy.

Figure 3 plots the statistical discrepancy using annualized seasonally adjusted quarterly data of the September 26th, 2013 vintage of BEA data. This data vintage was just after a BEA comprehensive revision, so the BEA revised each datapoint in Figure 3 at least once. Figure 3 reveals persistent differences between real GDP and real GDI even after a BEA comprehensive revision. The BEA's GDP figures are generally greater than its GDI figures until the mid-1990s, with real GDP exceeding real GDI by $250 billion (2009 chain-weighted dollars) in the first quarter of 1993. After the mid-1990s, GDI generally exceeds GDP up to a maximum of $259 billion (2009 chain-weighted dollars) in the third quarter of 2006.

The quarter-to-quarter variance of the statistical discrepancy has been widening over time, which may reflect the nonstationarity of real GDP and real GDI. Figure 4 plots the implied annual percent changes of the statistical discrepancy with the BEA's quarterly esti-

Fraumeni, 2008). Another reason why the BEA may prefer GDP to GDI is because the BEA publishes deflators for detailed components of GDP, but does not publish such deflators for GDI (Nalewaik, 2012).

mates. To reemphasize, the data in Figure 4 have been subject to at least one BEA comprehensive revision, with older data points undergoing multiple comprehensive revisions. From Figure 4, we can see considerable persistent differences in the quarter-to-quarter movements of GDP and GDI. From the 264 quarterly observations from the first quarter of 1947 to the second quarter of 2013, 58% have a statistical discrepancy of at least $\pm$ 1 percentage point, and 29% have a statistical discrepancy of at least $\pm$ 2 percentage points, with the mean magnitude of the discrepancy at 1.49 percentage points.

Some of the statistical discrepancy can be traced to the fact that, when constructing quarterly estimates of GDP or GDI, the BEA has some difficulty allocating output to a particular quarter.[11] Figure 5 plots 4-quarter percent changes of the statistical discrepancy. The smoothed 4-quarter percent changes of GDP and GDI have a smaller statistical discrepancy than the raw quarterly data, with the mean magnitude of the 4-quarter discrepancy at 0.59 percentage points. However, in some quarters, the magnitude of the 4-quarter discrepancy is still over 2 percentage points.[12]

3 Methodology

We defined our entire methodology prior to executing any analysis. Defining our methodology prior to the analysis carries at least four benefits: (1) we set a uniform standard for analyzing the results of models, (2) we avoid hindsight bias in model selection and analysis, (3) we avoid pretesting our results, and (4) we "tie our hands" to avoid both specification searching and p-hacking that would get us "stronger" (that is, potentially more publishable) results. In particular, because our methodology, like many other research studies, involves choices in situations where no dominant option exists, ***avoiding specification search-***

[11] To detect seasonality in the statistical discrepancy, we regress the quarterly statistical discrepancy on a full set of quarterly dummies, but this regression yields no statistically significant coefficients.

[12] Because of comprehensive revisions, which benchmark GDP and GDI to census data, the BEA will have relatively good estimates of the long-run levels of GDP and GDI. For example, the sample averages of the annual percent changes of quarterly real GDP and real GDI using the September 26th, 2013 vintage of data from the first quarter of 1947 to the second quarter of 2013 are both about 3.2%, implying a small statistical discrepancy in the long run.

ing —including specification searching disguised as "robustness checks" —is crucial for obtaining credible and transparent results. Examples of specification searching could include refitting the models in our sample using GDP or GDI data from different time periods, or fitting a variety of measurement error models until we found a particular model that gave statistically significant results for either GDP or GDI.

To analyze the effect of measurement error in US GDP on economic research, we start with a sample of 29 papers for which we were able to replicate the key published results using author-provided data and code files. Chang and Li (2015a, Forthcoming) provide full details of the replication sample, but the main characteristic of these 29 papers is that they all have key results that derive from GDP.[13] We identify the key results of each paper before reestimating any models, which requires some subjective decisions on our part. We attribute a key result of a paper to GDP when the authors themselves refer to GDP as driving a key result, or when a discussion of GDP is featured either in the abstract or prominently in the introduction of their work. Of course, it is possible that we misidentify some key results. But because we defined all key results prior to any reestimation, we hope that our procedure mitigates any bias that potential misidentification might cause.

The papers in our sample come from well-regarded, peer-reviewed economics journals: *American Economic Journal: Economic Policy*, *American Economic Journal: Macroeconomics*, *American Economic Review*, *Canadian Journal of Economics*, *Econometrica*, *Economic Journal*, *Journal of Applied Econometrics*, *Journal of Political Economy*, *Review of Economic Dynamics*, *Review of Economic Studies*, *Review of Economics and Statistics*, and the *Quarterly Journal of Economics*. Because these papers are from well-regarded, peer-reviewed journals, and because authors provide data and code files to run their models (either from journal replication archives, their personal websites, or to us through emails), we believe the quality and robustness of the research findings of these papers are very high.

[13] Our sampling frame also includes papers that use GDI as a key component of estimated models, but we were unable to locate any paper that uses GDI instead of GDP to estimate models. The dearth of papers that use GDI may be because the BEA features GDP more prominently than GDI in its press releases (Nalewaik, 2010). We do not take sides on whether GDP or GDI is a better indicator of unobserved output.

After replicating published results, we identify the original vintage of GDP that the published papers use by comparing the author-provided data files to historical BEA vintages of GDP. We also check the papers to see whether authors identify the original vintage of data they use. If these two procedures leave us unable to identify the original vintage and we have not contacted the authors requesting assistance with replication, then we email the authors about the original vintage of GDP they use, following the method in Chang and Li (2015a, Forthcoming). In most cases, we match the original vintage with this three-step procedure.[14] In some cases, a historical BEA vintage approximates the original vintage in the author-provided data files, but we do not find an exact match. For 3 of the 29 papers in our sample, we are unable to identify the original vintage of GDP used in the paper and hence exclude them from our analysis (Krishnamurthy and Vissing-Jorgensen, 2012; Mertens and Raven, 2011; Heutel, 2012).[15] We exclude two papers where we do not possess code to reestimate the models with alternative data (Schmitt-Grohé and Uribe, 2011, 2012).[16] We also exclude Clark and McCracken (2010) because the paper relies completely on real-time data that encompass many vintages of GDP, so we are unable to change a single original vintage for a current-vintage series.[17] Section 4 and the web appendix detail the original

[14]Because of different sample periods and the BEA's data revision schedule, on occasion we can match multiple vintages to the author's vintage. For example, suppose a paper estimates a model with data from 1984 Q1 to 2005 Q4 using the January 2007 vintage of GDP. Because the BEA only revises GDP more than one quarter back during an annual or comprehensive revision, the January 2007, February 2007, and March 2007 GDP vintages for 1984 Q1 to 2005 Q4 are all identical, as only 2006 Q4 GDP is different between these three vintages. When we are able to match more than one vintage, we use and report one of the observationally equivalent vintages in the web appendix available on Chang's website, https://sites.google.com/site/andrewchristopherchang/research.

[15]The most common cause of our inability to identify data vintages is because the author-provided data files only provide the transformed series used in the analysis and do not provide the raw data. For example, if GDP appears in the model as the debt-to-GDP ratio and the authors only include the debt-to-GDP ratio in the data file, then we cannot identify the GDP vintage.

[16]In this scenario, the original author-provided code files have parameter estimates hard-coded, which enables replication of the original tables and figures but does not allow for reestimation. When the original replication files lack code for reestimation, we email the authors requesting additional code to reestimate their models.

[17]An issue we do not investigate is the effect of using real-time vintages against end-of-sample vintages. Using real-time vintages instead of end-of-sample vintages may have implications for forecast accuracy (Koenig, Dolmas, and Piger, 2003; Chang and Hanson, 2016). The data we use in this paper are end-of-sample vintages.

vintages we identify for each paper in our sample.[18]

For our remaining sample of 23 papers, we reestimate the models but replace the original vintage of GDP with the original vintage of GDI, the current vintage of GDP, and the current vintage of GDI, where current vintage is the revised data as of September 26th, 2013.[19] When original-vintage GDP appears more than once in the estimated models, we replace original-vintage GDP wherever it appears.[20] For example, if a paper estimates a VAR with GDP and net exports where net exports is scaled by GDP, then we replace both the GDP variable and the denominator of the scaled net exports variable.

If the GDP deflator also appears in the estimated models, then when we reestimate the models using current-vintage data we also replace the original-vintage GDP deflator with the current-vintage GDP deflator. The BEA deflates GDP and GDI using the same GDP deflator, so our specifications with both current-vintage GDP and current-vintage GDI use the same vintage of the GDP deflator. We do not update the vintage of data other than the GDP deflator and GDP.[21]

Table 1 lists the papers in our analysis.[22]

4 Results

We find that using current-vintage GDP produces the same qualitative result as the original article for all 23 of our papers. For 3 of 23 papers, using either original-vintage GDI or current-vintage GDI instead of original-vintage GDP produces qualitatively different results

[18] The web appendix is available on Chang's website, https://sites.google.com/site/andrewchristopherchang/research.

[19] Because we want to focus on official statistics, we opt not to estimate models with other measures of latent output, such as the proposed filtered output measure by Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2016), or linear combinations of GDP and GDI.

[20] The Bureau of Economic Analysis (BEA) maintains quarterly US GDP data since 1947 and annual GDP data since 1929. If the paper uses a combination of pre-1947 and post-1947 quarterly GDP data, then we replace only data since 1947. Similarly, we only replace annual data since 1929.

[21] Following this definition, we do not update the vintage of other price deflators. For example, when papers deflate data with the core personal consumption expenditures price deflator, we do not update the vintage of this deflator.

[22] A researcher could characterize this study as the "scientific replication" of 23 papers, following the terminology of Hamermesh (2007).

than the original article. We focus on whether the qualitative conclusions change when estimating the original models with other measures of output for three reasons: (1) it is difficult to justify comparing quantitative differences across papers due to different models as papers report fundamentally different results, so quantitative comparisons are tenuous at best; (2) the policy recommendation of a paper would only substantively change when the fundamental qualitative conclusion of a paper is different; and (3) focusing on qualitative results also allows us to give a lower bound on the effect of measurement error of GDP on economic research, as we classify many quantitative differences as no qualitative change.

To give the reader an idea of how we classify results, this section first details a paper where we find results qualitatively similar to the original paper but where our quantitative estimates are different. This section then explains each of the papers where we find qualitatively different results after estimating the models using the other measures of unobserved output. The web appendix gives the results for the remaining papers, where we believe the results with the other measures of unobserved output are all qualitatively similar to the published results.

4.1 Auerbach and Gorodnichenko (2012, 2013)

We use Table 1 from Auerbach and Gorodnichenko (2012), as corrected in Auerbach and Gorodnichenko (2013) as an example of finding quantitatively different yet qualitatively similar results using our other measures of unobserved output. The web appendix shows our analysis of the other key figures from Auerbach and Gorodnichenko (2012).

Table 2 shows the published estimates of Table 1 from Auerbach and Gorodnichenko (2012) and Table 3 shows our replication results. Most of our replication estimates are within 10% of their reported values. We find a slightly higher defense spending multiplier in recessions (max multiplier of 4.27) than the authors do (max multiplier of 3.56). Our replication supports two of the main results of Auerbach and Gorodnichenko (2012): (1) higher fiscal multipliers in recessions than expansions and (2) large defense spending multipliers in recessions.

Table 4 shows our results from replacing original-vintage GDP with original-vintage GDI. With original-vintage GDI, we find a much higher defense spending multiplier in recessions (max multiplier of 6.15) and a defense spending multiplier for expansions that is always negative (max multiplier of -0.49). The estimate of the nondefense multiplier for recessions is also smaller (max multiplier of 0.51) than the published estimate (max multiplier of 1.22). In addition, the estimate of the government investment spending multiplier is almost zero for recessions (max multiplier of -0.08), whereas the published estimate is expansionary (max multiplier of 2.85). Nevertheless we continue to estimate higher fiscal multipliers in recessions than expansions for government consumption spending and total government spending, with government defense spending still having the highest multiplier. Therefore, we classify the results with original-vintage GDI as consistent with the published results. The web appendix shows our analysis of Auerbach and Gorodnichenko (2012) Table 1 with current-vintage GDP and current-vintage GDI, both of which give the same qualitative result as the published estimates.

We now turn to results where an alternative output measure gives different qualitative results than the published paper.

4.2 Corsetti, Meier, and Müller (2012)

Corsetti, Meier, and Müller (2012) explain their key empirical result as follows: an "increase in government spending causes a substantial rise in aggregate output... a positive spending shock triggers a sizable buildup of public debt, followed over time by a decline of government spending below trend" (pg. 878). The authors show these results from the impulse responses from vector autoregressions (VARs) in their Figures 1 and 2. Corsetti, Meier, and Müller (2012)'s Figure 1 identifies the VAR using the Blanchard and Perotti (2002) method, while Corsetti, Meier, and Müller (2012)'s Figure 2 identifies the VAR following Ramey (2011). Their measure of debt is the US debt-to-GDP ratio, so GDP appears twice in their baseline

VARs.[23] Our replication of these two figures, using data and code from the files posted at the *Review of Economics and Statistics*, match the published paper (Chang and Li, 2015a, Forthcoming).[24]

Figure 6 plots the impulse responses from the Corsetti, Meier, and Müller (2012) Figure 1 VAR using current-vintage GDP instead of original-vintage GDP as the measure of output. Figure 6 shows a statistically significant effect of government spending on output, with a multiplier of about 1. Debt-to-GDP continues to rise and then fall.

Figure 7 plots the impulse responses from the Corsetti, Meier, and Müller (2012) Figure 2 VAR using current-vintage GDP. As in Figure 6, output rises immediately following the government spending shock and the increase in output is statistically significant. The multiplier at the time of the government spending shock is, again, about 1. Debt-to-GDP rises and immediately falls.

Taken together, the evidence from Figures 6 and 7 are qualitatively consistent with the findings of Corsetti, Meier, and Müller (2012). Hence, we conclude that revisions to GDP have no qualitative effect on their results.

Figure 8 plots the impulse responses from the Corsetti, Meier, and Müller (2012) Figure 1 VAR using original-vintage GDI as the measure of unobserved output. Figure 8 shows similar debt-to-GDI dynamics as Corsetti, Meier, and Müller (2012), but the impulse response of GDI differs considerably from Corsetti, Meier, and Müller (2012). The effect of government spending on GDI immediately following the shock is no longer statistically significant and the point estimate of the multiplier is about zero. Further out, the effect of the government spending shock on GDI is negative and statistically significant about eight quarters following the shock.

Figure 9 plots the impulse responses from the Corsetti, Meier, and Müller (2012) Figure 2 VAR using original-vintage GDI. The figure continues to indicate that the government

[23] The Corsetti, Meier, and Müller (2012) specifications with net exports are also scaled by GDP so GDP appears three times in these VAR specifications, but their baseline VAR does not have net exports as a variable.

[24] We identify the Corsetti, Meier, and Müller (2012) GDP vintage as March 2010.

spending shock has no effect on GDI.

Figures 10 and 11 plot the Corsetti, Meier, and Müller (2012) impulse responses using current-vintage GDI. The results are similar to using original-vintage GDI: a government spending shock has a zero to negative effect on GDI.

Because using current-vintage GDP gives similar results to the original paper (a significant and positive government spending multiplier on output) and because both original-vintage GDI and current-vintage GDI indicate a zero or negative effect of government spending on output, we conclude that data revisions to the same measure of output have no qualitative effect on these results, but switching from GDP and GDI does qualitatively influence the results for this paper.

4.3 Inoue and Rossi (2011)

From the abstract of Inoue and Rossi (2011): "This paper investigates the sources of the substantial decrease in output growth volatility in the mid-1980s by identifying which of the structural parameters in a representative New Keynesian and structural VAR models changed." As highlighted in their introduction, Inoue and Rossi (2011) "focus on a representative New Keynesian model, although our main results are robust to standard VAR estimation as well as larger-scale DSGE model estimation" (pg. 1187).[25] The authors display their key results in their Tables 1 and 3. Inoue and Rossi (2011) Table 1 displays p-values for the hypothesis test of time-varying structural parameters in their representative New Keynesian model. Their null hypothesis is that the parameters are time-invariant, and they use the estimate of the set of stable parameters (ESS) procedure. Inoue and Rossi (2011) Table 3 lists the contributions to the variance of output, inflation, and the interest rate in their representative New Keynesian model, where each parameter is allowed to change from its estimated value during the Great Moderation to its estimated value pre-Great Moderation.

[25] We found the estimation results for Inoue and Rossi (2011) were slightly different between different versions of Matlab, but our qualitative conclusions for the effects of using different measures of output are robust to the version of Matlab we use.

Table 5 shows our replication of Inoue and Rossi (2011) Table 1. We continue to identify the volatility of the technology shock, σ_z, as the only parameter in their model that is constant over time. From Table 6, which shows our replication of Inoue and Rossi (2011)'s Table 3, the contributions to the change in implied volatility of output, inflation, and the interest rate from progressively letting parameters move from their Great Moderation values to their pre-Great Moderation values are all similar to their reported estimates.[26]

Table 7 shows Inoue and Rossi (2011)'s Table 1 reestimated with current-vintage GDP. The results show that the ESS procedure identifies the standard deviation of the cost-push shock, σ_e, as time-invariant in addition to σ_z. Table 8 shows Inoue and Rossi (2011)'s Table 3 reestimated with current-vintage GDP. While the contributions to the change in the implied volatility of output, inflation, and the interest rate are all a bit different from the published estimates and our replication results, the qualitative results continue to hold. The results from Table 8 indicate that progressively allowing parameters in the Inoue and Rossi (2011) New Keynesian model to be time-varying, according to the p-values of the Andrews (1993) Quandt Likelihood Ratio (QLR) stability test, implies that a time-varying standard deviation of the persistent monetary policy shock, σ_ν, and a time-varying persistence of the preference shock, ρ_a, would both significantly increase the volatility of output, inflation, and the interest rate. Allowing the standard deviation of the preference shock, σ_a, and the degree of inflation aversion of the Federal Reserve, ρ_π, to be time-varying would also have offsetting effects on volatility.

Table 9 shows Inoue and Rossi (2011) Table 1 reestimated with original-vintage GDI. The Inoue and Rossi (2011) ESS procedure now identifies two additional parameters, α and ψ, as time-invariant.

Table 10 shows Inoue and Rossi (2011) Table 3 reestimated with original-vintage GDI. The results of the table are qualitatively different than both the published results and the results estimated with current-vintage GDP. Focusing on the set of stable parameters, Table

[26]We find the Inoue and Rossi (2011) original GDP vintage is from August 2004.

10 shows that allowing σ_z to be time-varying would dampen output volatility in the estimated model as opposed to increasing output volatility as in Inoue and Rossi (2011). In addition, the reestimated contribution of σ_z to the volatility of inflation is over twice that of the published results. As far as the unstable parameters of the Inoue and Rossi (2011) model, the majority of the contributions to the volatilities of output, inflation, and the interest rate are much larger in magnitude and are frequently of opposite signs to the published results. For example, using original-vintage GDI causes us to estimate σ_a as dampening the volatilities of output and the interest rate by more than ten times the published estimates. The results with original GDI also show that σ_a has the effect of increasing the volatility of inflation, whereas the published estimate has σ_a as a negative contributor to the volatility of inflation.

Tables 11 and 12 show Inoue and Rossi (2011)'s Tables 1 and 3 reestimated with current-vintage GDI. The results are similar to the results with original-vintage GDI: the parameters have larger contributions, in magnitude, to the volatilities of output, inflation, and the interest rate that are frequency of the opposite sign as published estimates.

4.4 Morley and Piger (2012)

From the Morley and Piger (2012) abstract, the authors cite their key result as "...we construct a model-averaged measure of the business cycle. This measure also displays an asymmetric shape...", which is also consistent with the title of their paper, "The Asymmetric Business Cycle." The authors further elaborate on this result when they show their model-averaged measure of the business cycle in their Figure 3: "Perhaps the most striking feature of this [model-averaged] measure [of the business cycle] is its asymmetric shape, which it inherits from the bounceback models. In particular, the variation in the cycle is substantially larger during recessions than it is in expansions" (pg. 218). Our replication of this figure matches the result published in Morley and Piger (2012) and is shown in Figure 12.[27]

[27] We match the Morley and Piger (2012) GDP vintage to March 2007.

Figure 13 plots the Morley and Piger (2012) model-averaged measure of the business cycle estimated with current-vintage GDP, shown in their Figure 3. Figure 13 is qualitatively consistent with Morley and Piger (2012). The figure displays large dips in output during National Bureau of Economic Research (NBER) recessions, with a gradual run-up in output after the initial bounceback during NBER expansions.

Figure 14 plots the Morley and Piger (2012) model-averaged measure of the business cycle estimated with original-vintage GDI. This model-averaged measure displays much shallower recessions and larger run-ups in output just prior to a recession than the same measure estimated with either vintage of GDP. For example, during the tech bubble leading up to the 2001 recession, the Morley and Piger (2012) model-averaged measure of the business cycle estimated with original-vintage GDI more than doubles the measures estimated on either original-vintage GDP or current-vintage GDP. In the years leading up to the 1990 recession, the model-averaged measure of the business cycle estimated with original-vintage GDI exhibits much more volatility and a larger run-up prior to the 1990 recession than when the measure is estimated using either original-vintage GDP or current-vintage GDP.

Figure 15 plots the Morley and Piger (2012) model-averaged measure of the business cycle estimated with current-vintage GDI. The results are similar to when the measure is estimated with original-vintage GDI: larger run-ups in output prior to a recession and shallower recessions than when the measure is estimated with GDP.

Table 13 tests statistically the differences in the model-averaged measures of the business cycle. The table shows variances in NBER expansions and NBER recessions for the model-averaged measure estimated across original-vintage GDP, current-vintage GDP, original-vintage GDI, and current-vintage GDI, and the p-values from the F-test of equality of variance between expansions and recessions for each output estimate. For the two measures of the business cycle estimated with GDP, the variance of output in recessions is about twice that in expansions and the F-test rejects equality of variance between expansions and recessions at the 1% level, consistent with the findings of Morley and Piger (2012). For the model-averaged measure using original-vintage GDI, the variance of output in recessions is about 50% larger than the variance in expansions. The F-test for equality of variances is only marginally significant (p = 0.069). For the model-averaged measure using current-vintage GDI, the variance of output in recessions is only about 30% larger than the variance in expansions and the F-test is unable to reject equality of variances at standard levels (p = 0.199). We take this table as additional evidence that GDI may differ systematically from GDP due to measurement error and that the differences between GDP and GDI can influence published results.

5 Which Research Results Are More Fragile to Measurement Error?

In this section, we provide some speculation on what types of results are potentially more affected by measurement error. Although our sample of 23 papers is much larger than any other study that looks at the effect of measurement error on economics research, we do not have the statistical power to discriminate between many competing hypotheses on what characteristics of research are correlated with being more fragile to measurement error.[28]

The first natural question to ask is, what types of models may be more affected by measurement error? Table 14, Panel A splits our sample into papers with structural vs. reduced-form models. Of the three papers where we find different results when using GDI instead of GDP, Inoue and Rossi (2011) estimate a structural model while Corsetti, Meier, and Müller (2012) and Morley and Piger (2012) estimate reduced-form models. Therefore, we do not have strong evidence to suggest whether structural or reduced-form models may be more fragile.

As discussed in section 2, some of the statistical discrepancy can be traced to the fact

[28]In the literature on data revisions, Croushore and Stark (2002, 2003)'s sample of 3 papers is the largest sample that we are aware of.

that the BEA has some difficulty assigning output to a particular quarter. Table 14, Panel B splits our sample into papers that use annual vs. quarterly data. All three papers where we find different results using GDI instead of GDP estimate models with quarterly data. Therefore, estimating models with smoothed annual data may mitigate measurement error issues, but the tradeoff is a reduced sample size.

6 Conclusion

We investigate the effect that measurement error in latent US output has on economic research using two approaches. First, we use data revisions to GDP, which is the BEA's estimate of latent US output based on expenditure data. Second, we use GDI, a theoretically identical estimate of latent US output that the BEA creates with income data. To our knowledge, this paper is the first to use the fact that a national statistical agency produces two theoretically identical estimates of the same unobserved variable with independent data to look at the effect that measurement error has on economic research. Existing studies that look at this effect only use data revisions.

Using a sample of 23 published economics articles from well-regarded peer-reviewed journals, we find that revisions to GDP have no qualitative effect on published results. However, for 3 of 23 papers, estimating models with GDI changes their qualitative conclusions.

Our result that revisions to GDP have no effect on published research is at odds with the literature that we are aware of, which generally concludes that data revisions have an effect. For example, Croushore and Stark (2002, 2003) find that using revised data qualitatively alters the results of Hall (1978) and Blanchard and Quah (1989), although they find no effect of data revisions on Kydland and Prescott (1990). Ponomareva and Katayama (2010) compare using different vintages of the Penn World Tables (PWT) on the conclusions of Ramey and Ramey (1995) and find that newer versions of the PWT alter the original results. Faust, Rogers, and Wright (2003) re-run the model of Mark (1995) using successive data vintages up to October 2000 and find that newer data vintages generate results that are at odds with Mark (1995).

We outline two reasons why we believe our finding that revisions to data have no effect on published results may be different than the literature.

The first reason is that the time dimension of our data revisions is a bit shorter than the literature. The median paper in our sample uses a GDP vintage from July 2008. Therefore, the median time gap from original vintage to current vintage is 5 years and 2 months. Croushore and Stark (2002, 2003) update the original vintage of Hall (1978) (original-vintage May 1977) and Blanchard and Quah (1989) (original-vintage February 1988) to a current vintage of February 1998. Ramey and Ramey (1995) use PWT 5.0 (May 1991), and Ponomareva and Katayama (2010) compare that version to PWT 6.1 (October 2002). Faust, Rogers, and Wright (2003) use successive vintages from Mark (1995)'s original vintage of April 1992 up to October 2000, although they find that data vintages a mere two years away from Mark (1995)'s original vintage generate qualitatively different conclusions.[29]

The second reason for why we find a different effect of data revisions than the literature does may be that existing studies select either a single paper or select a small sample of papers to illustrate their claims. Because of the potential editorial preference for significant results, it is possible that the papers we are aware of (and cite in this article) are biased toward finding significant results, which would be an illustration of the Rosenthal (1979) file-drawer problem where papers that find insignificant results are locked away in file drawers and never published. Because our sample of papers spans multiple journals across topic areas in macroeconomics, we feel our result that data revisions do not qualitatively affect economics research is less likely to suffer from the file-drawer problem.[30]

[29]A potential suggestion that an editor or referee may make to us in the future is to reestimate the models in our sample using even newer data, as our current vintage of data is from September 26th, 2013. We are against this idea. We (and the editor or referee who would make this suggestion) have already observed the results using the September 26th, 2013 vintage of data and, should we reestimate the models using newer data, the estimation would have been conditioned on observing the results with the September 26th, 2013 vintage of data and would be pretested. Similarly, reestimating the models in our sample using a linear combination of GDP and GDI would also yield results that would be pretested.

[30]Brodeur, Lé, Sangnier, and Zylberberg (2016) provide evidence on the Rosenthal (1979) file-drawer

Our finding that results from models estimated with current-vintage GDP can differ from models estimated with current-vintage GDI supports the hypothesis that measurement error in the National Income and Product Accounts (NIPAs) does not revise away with multiple data revisions. Because the difference between original-vintage and current-vintage always spans at least one BEA comprehensive revision, we find that measurement error in the NIPAs does not revise away even after a BEA comprehensive revision, consistent with research by Nalewaik (2010, 2014).

We assert that measurement error in macroeconomic data can have meaningful consequences on research because we find that estimating models using GDI instead of GDP can change published results. We recommend that economic models should take into account when data are the estimates of the true quantities of interest. For the specific context of models estimated with GDP, we suggest that estimation should be robust to using either GDP or GDI as an author's estimate of latent output. In other situations where multiple estimates of the identical quantity of interest are available, such as with balance of payments data, results should be robust to using multiple estimates.[31] More generally, when the quantity of interest is a latent variable and not the statistic that estimates the latent variable, research results should also be robust to using data with different definitions of the latent variable. For example, when estimating a model on US data that includes inflation, the results should be robust to using different measures of US inflation across data definitions, such as: consumer price index inflation, core personal consumption expenditures (PCE) inflation, total PCE inflation, etc.

An assumption behind our assertion on measurement error's effect on research is that latent output is the object of interest behind the papers in our sample. This assumption could fail if authors estimate models that account for the measurement error that is only in GDP and is not in GDI. We are not aware of any research into the GDP or GDI statistics

problem in top economics journals.

[31] Similar to its GDP and GDI statistics, the BEA's estimate of the US current account should, in theory, equal the US capital account. Because of measurement error these quantities also differ in practice.

that is able to differentiate the measurement error in the two statistics to such a fine degree. Nalewaik (2010, 2014)'s research into the GDP and GDI statistics concludes that there is both classical measurement error and a loss of signal measurement error in both GDP and GDI, but it does not isolate a source or form of measurement error that is specific to GDP and not to GDI. We also do not believe the papers in our study differentiate the measurement error specific to GDP that is not present in GDI. Most papers in our sample ignore measurement error.

Another reason why latent output may not be the object of interest is if authors conduct research into the national statistics themselves instead of the objects the statistics estimate, such as by looking into the effects of macroeconomic data announcements on the stock market or foreign exchange rates (e.g., Faust, Rogers, Wang, and Wright, 2007; Rangel, 2011). From our reading of the papers where we find significantly different results than the authors using GDI, we believe the object of interest of the papers is latent output, not the GDP statistic.[32]

Overall, we view our results as a lower bound on the potential effect that measurement error in macroeconomic data has on economic research because of three factors.

First, we draw our sample of papers only from published research in well-regarded journals. These papers all survived intense peer review, which includes a barrage of reported robustness checks and, presumably, another barrage of unreported robustness checks that confirm the published findings.[33]

[32]For Corsetti, Meier, and Müller (2012), the authors describe their result as follows: an "increase in government spending causes a substantial rise in aggregate output... a positive spending shock triggers a sizable buildup of public debt, followed over time by a decline of government spending below trend" (pg. 878). Corsetti, Meier, and Müller (2012) do not reference GDP until section 2. According to its abstract, Inoue and Rossi (2011) "investigates the sources of the substantial decrease in output growth volatility in the mid-1980s" – instead of, perhaps, investigating the source of the substantial decrease in the volatility of the GDP statistic in the mid-1980s. Their results are also framed in terms of output, not GDP. In addition, Inoue and Rossi (2011) do not mention GDP until the third section (methodology). Similarly, Morley and Piger (2012), in their analysis of the asymmetric business cycle, argue that "the model averaged measure of the business cycle captures a meaningful macroeconomic phenomenon and sheds more light on the nature of fluctuations in aggregate economic activity than *simply looking at the level or the growth rates of real GDP*" (pg. 208-209, emphasis added). That is, Morley and Piger (2012) are interested in general real business cycle patterns, not the pattern of the GDP statistic, and their academic contribution is to improve on just looking at GDP.

[33]You could imagine a scenario where unpublished working papers have more robust results than published papers, but that scenario would be particularly discouraging for maintaining publication as an outlet for

Second, in our exercise we only affect the GDP series used in the original paper by updating the vintage to current vintage, switching GDP to GDI, or both. In models that use multiple data series, we leave the remainder of the data the same as in the published work. If we were to modify all variables included in multivariate models, then the potential effect of measurement error across all variables could be greater than simply the measurement error in GDP.[34]

Third, although the BEA does not publish estimates of the measurement error in its GDP or GDI statistics, we conjecture that, because the BEA devotes considerable time and effort to measuring GDP and GDI and has no direct incentive to misreport either GDP or GDI, these statistics are well-measured relative to other macroeconomic data. Data from countries that have far fewer resources to devote to their national accounts or data from countries that have incentives to misreport may contain more measurement error than US GDP or GDI.[35] Because we focus on research papers that use potentially well-measured estimates of US output, results from other studies that rely on other macroeconomic data could be more fragile than those in our sample.

A limitation of our analysis is that we do not discern which measure of output, GDP or GDI, is closer to true unobserved output.[36] The purpose of this paper is to show that measurement error in national statistics could affect economics research and we have provided a broad scope of examples to this effect.

scholarly communication.

[34] A researcher may be concerned that, because we only update the vintage of GDP, GDI, and the deflator, that using current-vintage output data and original-vintage data for non-output variables may disrupt the sample moments (for example, covariances) between output and the non-output variables, which would lead our study to give erroneous results if the BEA revises all of its data to keep the original sample moments intact. As we find that using current-vintage output data and original-vintage non-output data give similar results as the original published works, which use original-vintage data, we feel that our results are robust to this concern. Rather, we find that using GDI instead of GDP irrespective of the vintage of GDI can give meaningful differences in model output, which cannot be explained merely by the sample moments between output and non-output variables in original-vintage data.

[35] Michalski and Stoltz (2013) find evidence that countries may strategically misreport their economic data.

[36] For a discussion on this issue, see Nalewaik (2010, 2014) who asserts that GDI is superior to GDP. See also comments on Nalewaik (2010) by Diebold (2010) and Landefeld (2010), as well as work by Fleischman and Roberts (2011).

Table 1: Papers Under Study

Paper
Auerbach and Gorodnichenko (2012, 2013)
Barro and Redlick (2011)
Baumeister and Peersman (2013)
Canova and Gambetti (2010)
Carey and Shore (2013)
Chen, Curdia, and Ferrero (2012)
Corsetti, Meier, and Müller (2012)
D'Agostino and Surico (2012)
Den Haan and Sterk (2011)
Favero and Giavazzi (2012)
Gabaix (2011)
Hansen, Lunde, and Nason (2011)
Inoue and Rossi (2011)
Ireland (2009)
Kilian (2009)
Kormilitsina (2011)
Mavroeidis (2010)
Mertens and Ravn (2013)
Morley and Piger (2012)
Nakov and Pescatori (2010)
Ramey (2011)
Reis and Watson (2010)
Romer and Romer (2010)

Table 2: Auerbach and Gorodnichenko (2012) Table 1, Top Panel Published Results

	Max Point Estimate	Standard Error	Cumulative Point Estimate	Standard Error
Total Spending				
Linear	1.00	0.32	0.57	0.25
Expansion	0.57	0.12	-0.33	0.2
Recession	2.48	0.28	2.24	0.24
Defense Spending				
Linear	1.16	0.52	-0.21	0.27
Expansion	0.8	0.22	-0.43	0.24
Recession	3.56	0.74	1.67	0.72
Nondefense Spending				
Linear	1.17	0.19	1.58	0.18
Expansion	1.26	0.14	1.03	0.15
Recession	1.12	0.27	1.09	0.31
Consumption Spending				
Linear	1.21	0.27	1.2	0.31
Expansion	0.17	0.13	-0.25	0.1
Recession	2.11	0.54	1.47	0.31
Investment Spending				
Linear	2.12	0.68	2.39	0.67
Expansion	3.02	0.25	2.27	0.15
Recession	2.85	0.36	3.42	0.38

Corrected results from Auerbach and Gorodnichenko (2013). Table shows output multipliers for a $1 increase in government spending.

Table 3: Auerbach and Gorodnichenko (2012) Table 1, Top Panel With Original-Vintage GDP (Replication)

	Max Point Estimate	Standard Error	Cumulative Point Estimate	Standard Error
Total Spending				
Linear	0.89	0.29	0.60	0.23
Expansion	0.49	0.13	-0.80	0.16
Recession	2.12	0.18	2.17	0.19
Defense Spending				
Linear	1.53	0.56	0.39	0.22
Expansion	0.76	0.21	-0.94	0.26
Recession	4.27	0.93	2.18	0.78
Nondefense Spending				
Linear	1.69	0.08	2.09	0.15
Expansion	1.20	0.16	1.16	0.15
Recession	1.06	0.30	1.10	0.32
Consumption Spending				
Linear	0.83	0.28	0.90	0.29
Expansion	0.10	0.12	-0.16	0.12
Recession	2.16	0.65	1.33	0.36
Investment Spending				
Linear	2.06	0.60	2.75	0.60
Expansion	2.86	0.27	2.03	0.17
Recession	2.79	0.53	4.18	0.46

Replication of Table 1 of Auerbach and Gorodnichenko (2012) as corrected in Auerbach and Gorodnichenko (2013). Source: Chang and Li (2015a, Forthcoming). Table shows output multipliers for a $1 increase in government spending.

Table 4: Auerbach and Gorodnichenko (2012) Table 1, Top Panel With Original-Vintage GDI

	Max Point Estimate	Standard Error	Cumulative Point Estimate	Standard Error
Total Spending				
Linear	0.14	0.22	-0.03	0.24
Expansion	0.10	0.15	-1.68	0.20
Recession	1.18	0.16	1.38	0.17
Defense Spending				
Linear	0.42	0.23	-0.09	0.26
Expansion	-0.49	0.24	-3.05	0.37
Recession	6.15	0.84	2.20	0.65
Nondefense Spending				
Linear	1.86	0.08	2.03	0.17
Expansion	1.13	0.22	0.82	0.19
Recession	0.51	0.26	0.46	0.27
Consumption Spending				
Linear	0.54	0.25	0.36	0.28
Expansion	-0.06	0.13	-0.75	0.16
Recession	3.06	0.69	1.91	0.42
Investment Spending				
Linear	0.94	0.52	0.62	0.59
Expansion	3.11	0.29	3.02	0.24
Recession	-0.08	0.84	-1.95	0.55

Table shows output multipliers for a $1 increase in government spending.

Table 5: Inoue and Rossi (2011) Table 1 With Original-Vintage GDP (Replication)

Model Parameters	Individual p-Value	ESS p-Value
ρ_e	0	0
σ_ν	0	0
α	0	0
σ_a	0	0
σ_π	0	0
ρ_a	0	0
γ	0	0
ψ	0	0.01
ρ_{gy}	0	0
σ_e	0	0
ρ_υ	0	0
ρ_π	0	0
σ_z	1	1

Original GDP vintage is August 27, 2004. Set of stable parameters (90% probability level): S = $\{\sigma_z\}$. This table reports p-values of the QLR stability test (Andrews, 1993) on individual parameters, labeled "Individual p-value," and the p-values of each step of the Inoue and Rossi (2011) ESS procedure, labeled "ESS p-value." Source: Chang and Li (2015a, Forthcoming).

Table 6: Inoue and Rossi (2011) Table 3 With Original-Vintage GDP (Replication)

Parameter:	Output	Inflation	Interest Rate
No change: (actual S.D.)	0.89	0.48	0.30
Unstable Parameters:	% Contribution to Change		
ρ_e	7%	10%	-1%
σ_ν	71%	35%	40%
α	-2%	12%	1%
σ_a	-22%	-4%	-104%
σ_π	4%	15%	35%
ρ_a	25%	2%	94%
γ	20%	0%	18%
ψ	0%	0%	0%
ρ_{gy}	-43%	1%	24%
σ_e	-2%	-5%	-1%
ρ_υ	6%	5%	-15%
ρ_π	-13%	-23%	5%
Stable Parameters:			
σ_z	49%	53%	3%
All change: (actual S.D.)	1.45	0.92	0.39

Original GDP vintage is August 27, 2004. Set of stable parameters (90% probability level): S = $\{\sigma_z\}$. This table shows the percentage contribution to the increase or decrease in the volatilities of output, inflation, and the interest rate by progressively allowing each parameter to be time-varying, ordered according to the p-values of the QLR stability test (Andrews, 1993). Source: Chang and Li (2015a, Forthcoming).

Table 7: Inoue and Rossi (2011) Table 1 With Current-Vintage GDP

Model Parameters	Individual p-Value	ESS p-Value
ρ_e	0	0
σ_ν	0	0
α	0	0
σ_a	0	0
σ_π	0	0
ρ_a	0	0
γ	0	0
ψ	0	0
ρ_{gy}	0	0
σ_e	1	1
ρ_υ	0	0
ρ_π	0	0
σ_z	1	1

Set of stable parameters (90% probability level): S = $\{\sigma_e, \sigma_z\}$. This table reports p-values of the QLR stability test (Andrews, 1993) on individual parameters, labeled "Individual p-value," and the p-values of each step of the Inoue and Rossi (2011) ESS procedure, labeled "ESS p-value."

Table 8: Inoue and Rossi (2011) Table 3 With Current-Vintage GDP

Parameter:	Output	Inflation	Interest Rate
No change: (actual S.D.)	0.92	0.49	0.30
Unstable Parameters:	% Contribution to Change		
ρ_e	5%	7%	0%
σ_ν	98%	48%	94%
α	-2%	9%	2%
σ_a	-33%	-6%	-96%
σ_π	4%	10%	17%
ρ_a	23%	2%	67%
γ	34%	1%	1%
ψ	0	0	0
ρ_{gy}	-72%	4%	19%
ρ_υ	8%	5%	-11%
ρ_π	-15%	-29%	6%
Stable Parameters:			
σ_e	-1%	-1%	0%
σ_z	50%	50%	1%
All change: (actual S.D.)	1.38	0.90	0.38

Set of stable parameters (90% probability level): S = $\{\sigma_e, \sigma_z\}$. This table shows the percentage contribution to the increase or decrease in the volatilities of output, inflation, and the interest rate by progressively allowing each parameter to be time-varying, ordered according to the p-values of the QLR stability test (Andrews, 1993).

Table 9: Inoue and Rossi (2011) Table 1 With Original-Vintage GDI

Model Parameters	Individual p-Value	ESS p-Value
ρ_e	0	0
σ_ν	0	0
α	1	1
σ_a	0	0
σ_π	0.02	0
ρ_a	0	0
γ	0	0
ψ	0.09	0.19
ρ_{gy}	0	0
σ_e	1	1
ρ_υ	0	0
ρ_π	0	0
σ_z	1	1

Original GDI vintage is August 27, 2004. Set of stable parameters (90% probability level): S = $\{\alpha, \sigma_e, \sigma_z, \psi\}$. This table reports p-values of the QLR stability test (Andrews, 1993) on individual parameters, labeled "Individual p-value," and the p-values of each step of the Inoue and Rossi (2011) ESS procedure, labeled "ESS p-value."

Table 10: Inoue and Rossi (2011) Table 3 With Original-Vintage GDI

Parameter:	Output	Inflation	Interest Rate
No change: (actual S.D.)	0.98	0.93	0.35
Unstable Parameters:	% Contribution to Change		
ρ_e	2%	-19%	0%
σ_ν	50%	-192%	40%
σ_a	-346%	78%	-1663%
σ_π	1%	-16%	17%
ρ_a	-52%	3%	-292%
γ	690%	-189%	988%
ρ_{gy}	-193%	-355%	964%
ρ_υ	5%	-13%	-27%
ρ_π	-12%	681%	75%
Stable Parameters:			
α	0%	0%	0%
σ_e	0%	-3%	0%
σ_z	-45%	126%	-2%
ψ	0%	0%	0%
All change: (actual S.D.)	1.31	0.84	0.39

Original GDI vintage is August 27, 2004. Set of stable parameters (90% probability level): S = $\{\alpha, \sigma_e, \sigma_z, \psi\}$. This table shows the percentage contribution to the increase or decrease in the volatilities of output, inflation, and the interest rate by progressively allowing each parameter to be time-varying, ordered according to the p-values of the QLR stability test (Andrews, 1993).

Table 11: Inoue and Rossi (2011) Table 1 With Current-Vintage GDI

Model Parameters	Individual p-Value	ESS p-Value
ρ_e	0	0
σ_ν	0	0
α	0	0
σ_a	0	0
σ_π	0	0
ρ_a	0	0
γ	0	0
ψ	0	0
ρ_{gy}	1	1
σ_e	0.19	0.07
ρ_υ	0	0
ρ_π	0.04	0
σ_z	0.71	0.75

Set of stable parameters (90% probability level): S = $\{\rho_{gy}, \sigma_z\}$. This table reports p-values of the QLR stability test (Andrews, 1993) on individual parameters, labeled "Individual p-value," and the p-values of each step of the Inoue and Rossi (2011) ESS procedure, labeled "ESS p-value."

Table 12: Inoue and Rossi (2011) Table 3 With Current-Vintage GDI

Parameter:	Output	Inflation	Interest Rate
No change: (actual S.D.)	0.89	0.51	0.29
Unstable Parameters:	% Contribution to Change		
ρ_e	2%	5%	0%
σ_ν	2%	7%	22%
α	0%	5%	0%
σ_a	-47%	-3%	-93%
σ_π	0%	6%	3%
ρ_a	173%	5%	332%
γ	-97%	-3%	-140%
ψ	0%	0%	0%
σ_e	0%	-1%	0%
ρ_υ	1%	3%	-9%
ρ_π	1%	11%	-11%
Stable Parameters:			
ρ_{gy}	2%	-3%	-5%
σ_z	64%	67%	1%
All change: (actual S.D.)	1.49	1.14	0.39

Set of stable parameters (90% probability level): S = $\{\rho_{gy}, \sigma_z\}$. This table shows the percentage contribution to the increase or decrease in the volatilities of output, inflation, and the interest rate by progressively allowing each parameter to be time-varying, ordered according to the p-values of the QLR stability test (Andrews, 1993).

Table 13: Morley and Piger (2012) Model-Averaged Measure Variances

	Models Estimated With:			
	Original-Vintage GDP (Replication)	Current-Vintage GDP	Original-Vintage GDI	Current-Vintage GDI
Variance of NBER Expansions	0.340	0.337	0.310	0.314
Variance of NBER Recessions	0.631	0.642	0.463	0.417
F-test (p-value)	0.005	0.003	0.069	0.199

We calculate variances based on Morley and Piger (2012)'s model-averaged measure of the business cycle. The replication and original-vintage GDI columns use revised data as of March 30, 2007. Current-vintage data columns use revised data as of September 26th, 2013. F-tests for the equality of variances between National Bureau of Economic Research (NBER) expansions and NBER recessions for each model-averaged measure, H_0 : variances are equal, H_A : variances are different.

Table 14: Characteristics of Sample

Panel A: Model Type	
Structural	Reduced-Form
Chen, Curdia, and Ferrero (2012)	Auerbach and Gorodnichenko (2012, 2013)
Inoue and Rossi (2011)	Barro and Redlick (2011)
Kormilitsina (2011)	Baumeister and Peersman (2013)
Mavroeidis (2010)	Canova and Gambetti (2010)
Nakov and Pescatori (2010)	Carey and Shore (2013)
	Corsetti, Meier, and Müller (2012)
	Den Haan and Sterk (2011)
	D'Agostino and Surico (2012)
	Favero and Giavazzi (2012)
	Gabaix (2011)
	Hansen, Lunde, and Nason (2011)
	Ireland (2009)
	Mertens and Ravn (2013)
	Morley and Piger (2012)
	Ramey (2011)
	Reis and Watson (2010)
	Romer and Romer (2010)

Panel B: Data Frequency	
Annual	Quarterly
Barro and Redlick (2011)	Auerbach and Gorodnichenko (2012, 2013)
Canova and Gambetti (2010)	Baumeister and Peersman (2013)
Carey and Shore (2013)	Chen, Curdia, and Ferrero (2012)
Gabaix (2011)	Corsetti, Meier, and Müller (2012)
	Den Haan and Sterk (2011)
	D'Agostino and Surico (2012)
	Favero and Giavazzi (2012)
	Hansen, Lunde, and Nason (2011)
	Inoue and Rossi (2011)
	Ireland (2009)
	Kilian (2009)
	Kormilitsina (2011)
	Mavroeidis (2010)
	Mertens and Ravn (2013)
	Morley and Piger (2012)
	Nakov and Pescatori (2010)
	Ramey (2011)
	Reis and Watson (2010)
	Romer and Romer (2010)

Panel A classifies the sample into structural and reduced-form papers, where the reduced-form category includes any vector autoregressions. Panel B separates papers into estimation with annual or quarterly data. In the annual data column, Canova and Gambetti (2010) use 4-quarter changes and Carey and Shore (2013) uses the sum of several previous year's GDP.

Figure 1: GDP Revisions from September 2008 to September 2013 - Nominal Levels

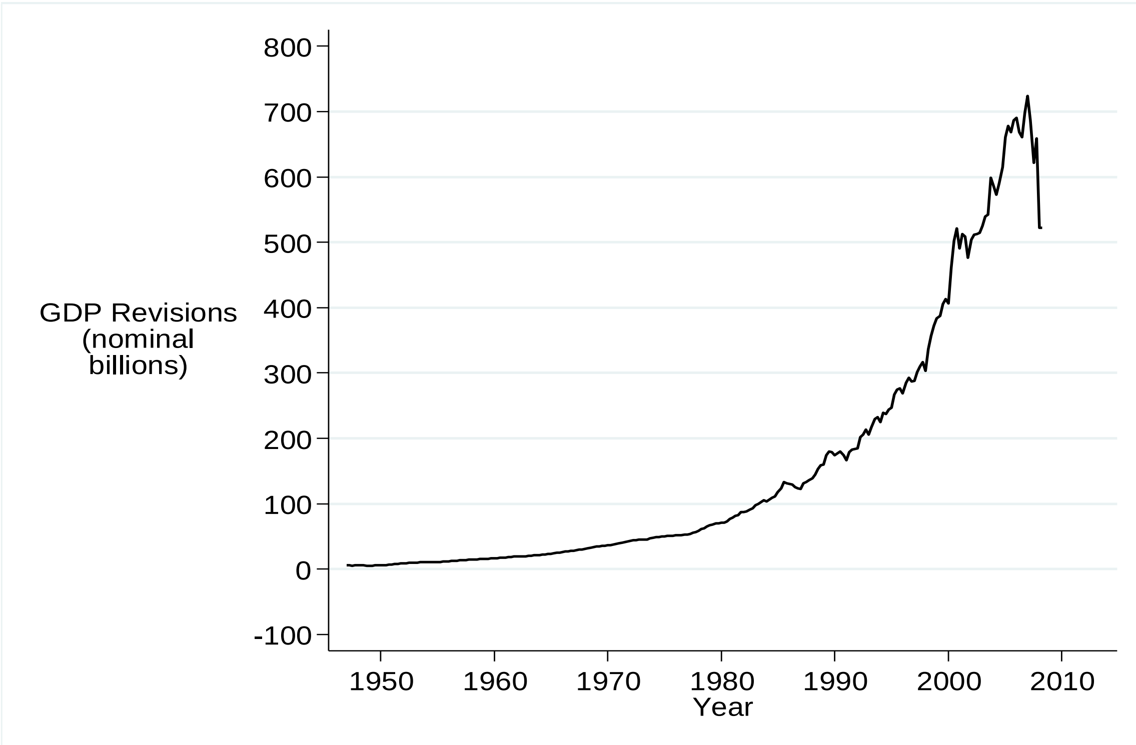


Figure plots annualized, seasonally adjusted, quarterly nominal GDP from the September 26th, 2013 vintage minus annualized, seasonally adjusted, quarterly nominal GDP from the September 26th, 2008 vintage.

Figure 2: GDP Revisions from September 2008 to September 2013 - Annual Percent Changes

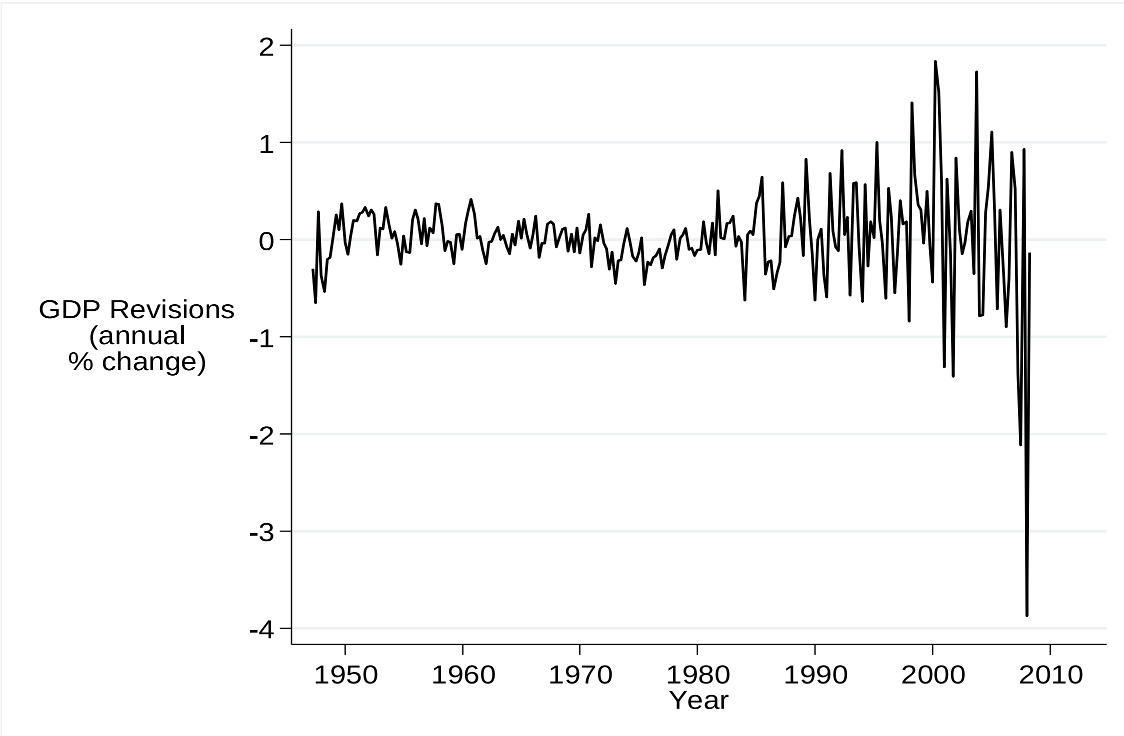


Figure plots annual percent changes of seasonally adjusted quarterly nominal GDP from the September 26th, 2013 vintage minus annual percent changes of seasonally adjusted quarterly nominal GDP from the September 26th, 2008 vintage.

Figure 3: The Statistical Discrepancy - Real Levels

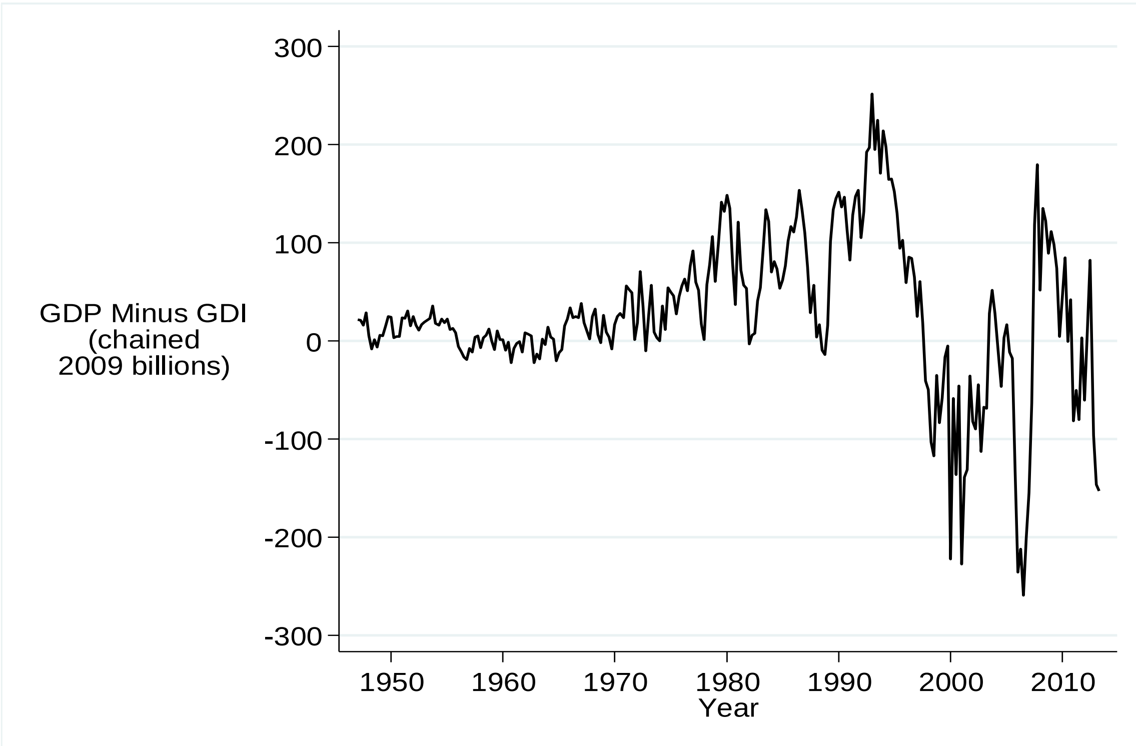


Figure plots annualized, seasonally adjusted, quarterly real GDP minus annualized, seasonally adjusted, quarterly real GDI using the September 26th, 2013 vintage of BEA data.

Figure 4: The Statistical Discrepancy - Quarterly Data in Real Annual Percent Changes

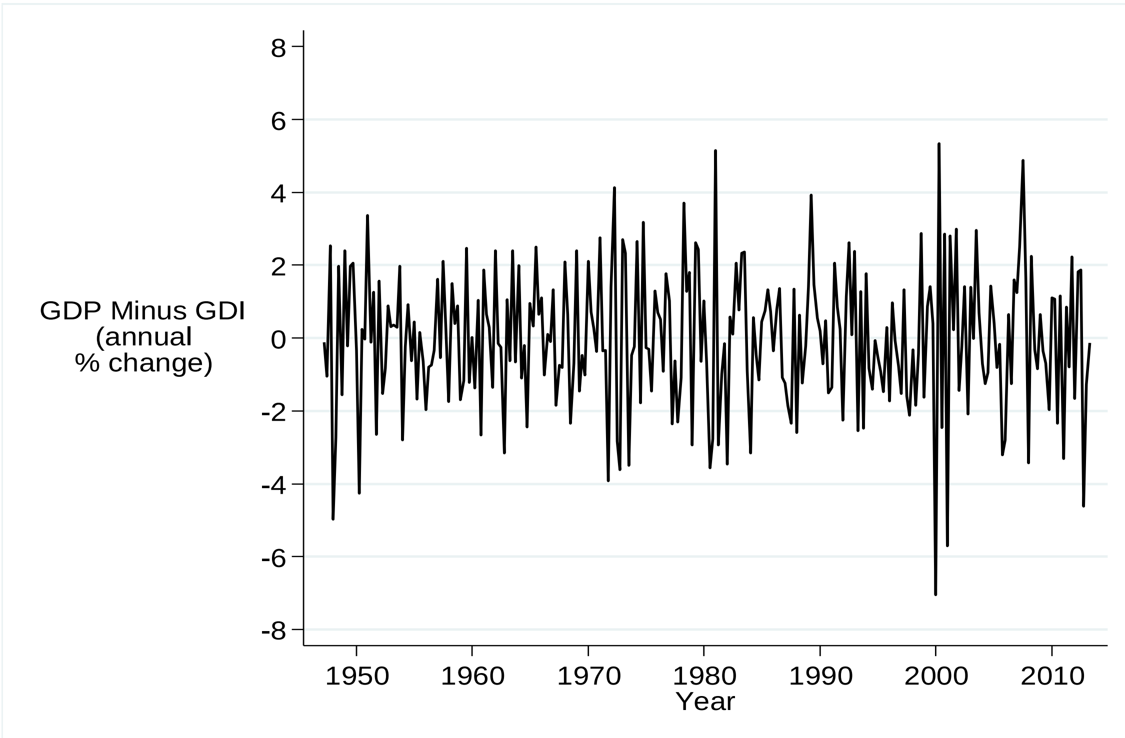


Figure plots annual percent changes of seasonally adjusted quarterly real GDP minus annual percent changes of seasonally adjusted quarterly real GDI using the September 26th, 2013 vintage of BEA data.

Figure 5: The Statistical Discrepancy - 4 Quarter Percent Changes

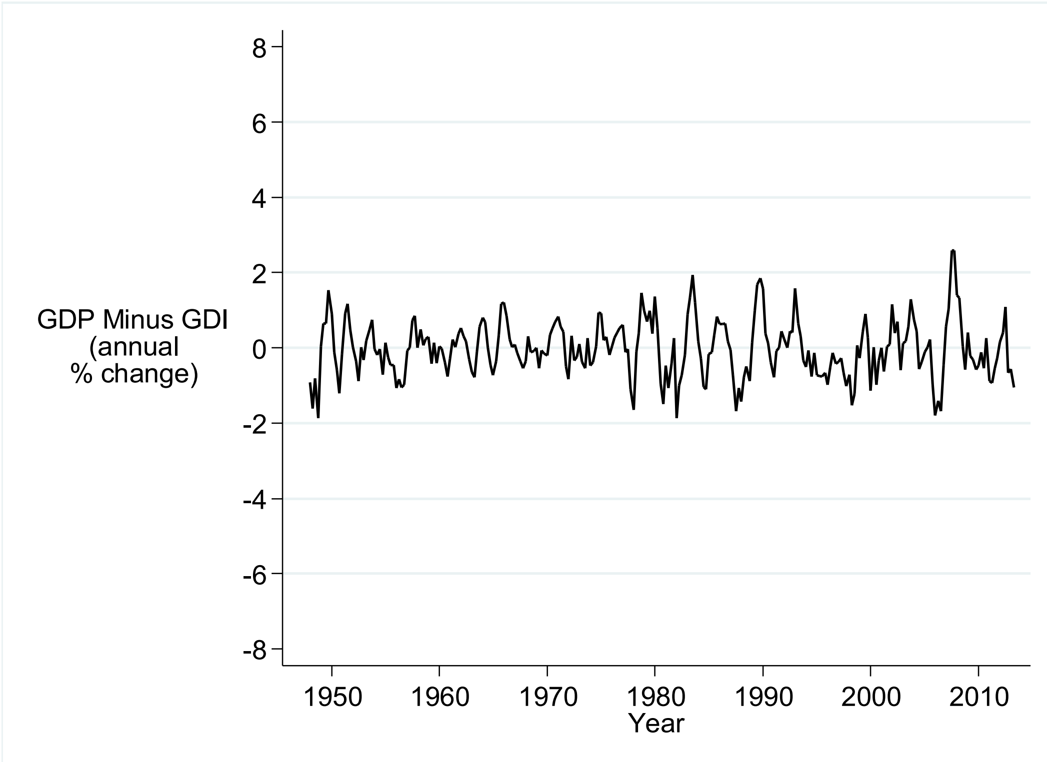
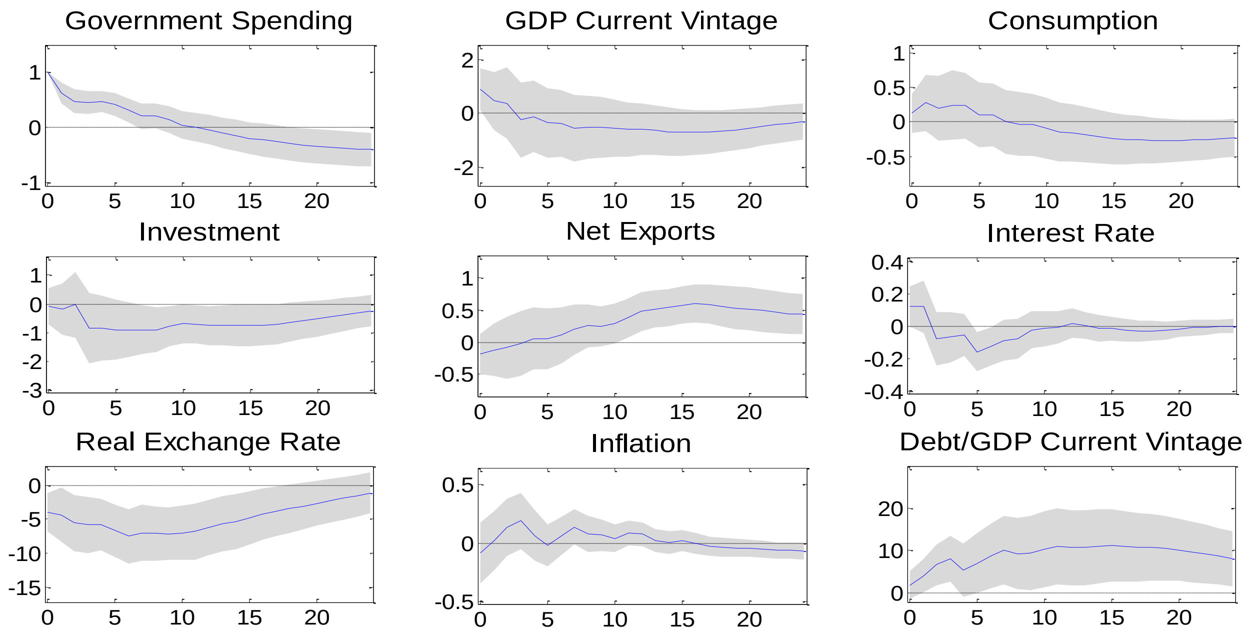


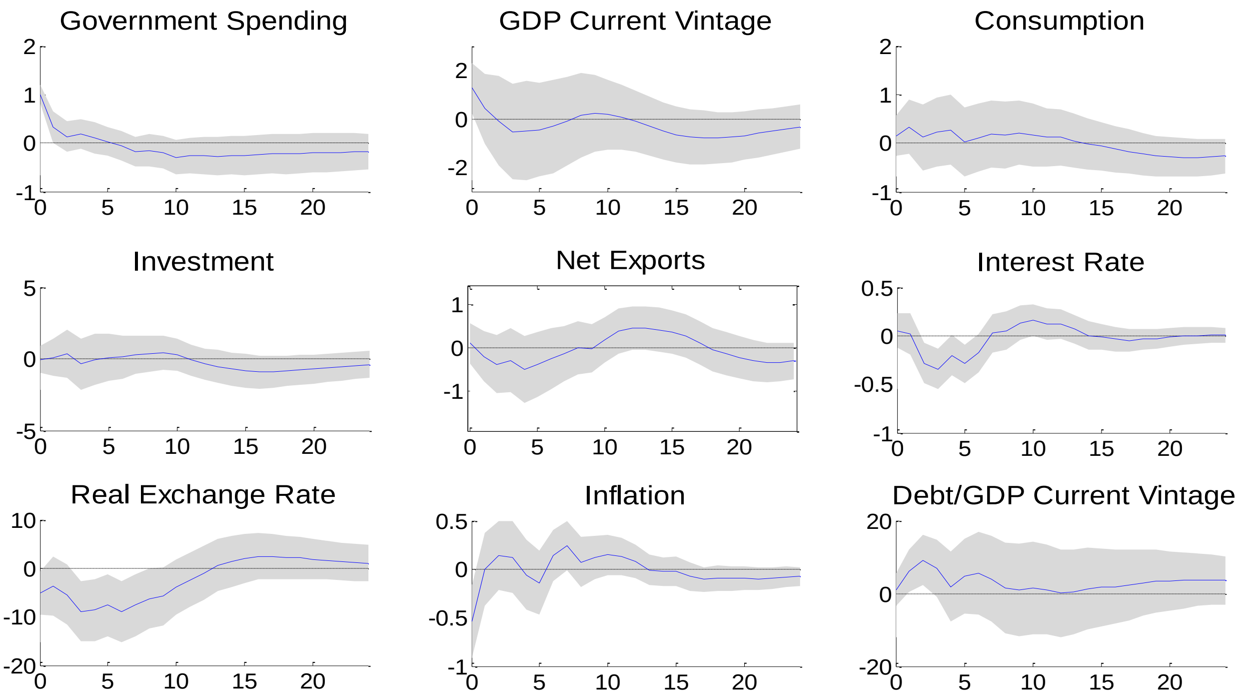
Figure plots 4-quarter percent changes of seasonally adjusted real GDP minus 4-quarter percent changes of seasonally adjusted real GDI using the September 26th, 2013 vintage of BEA data.

Figure 6: Corsetti, Meier, and Müller (2012) Figure 1 With Current-Vintage GDP



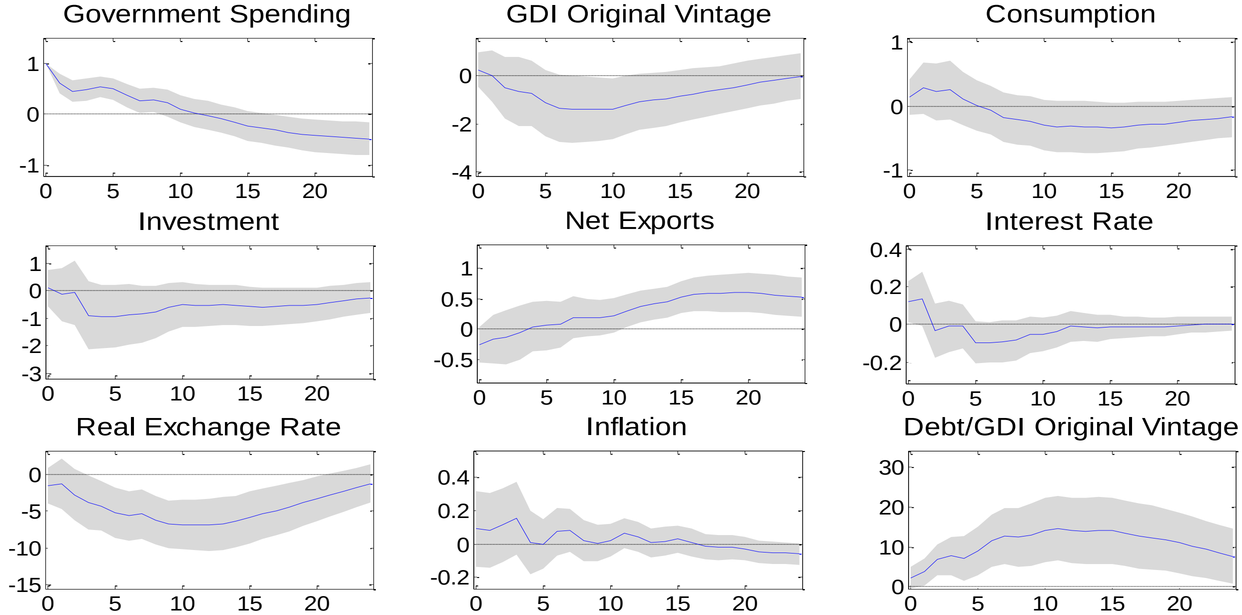
Impulse responses from a vector autoregression identified with the Blanchard and Perotti (2002) method. Solid blue lines indicate the point estimate. Grey area indicates the 90% confidence interval. Horizontal axis indicates quarters. Vertical axes denotes deviations from trend in percent points of trend output (in the case of quantities); percentage deviations from the preshock level (real exchange rate); and deviations from the preshock level in terms of quarterly percentage points (real interest rate and inflation).

Figure 7: Corsetti, Meier, and Müller (2012) Figure 2 With Current-Vintage GDP



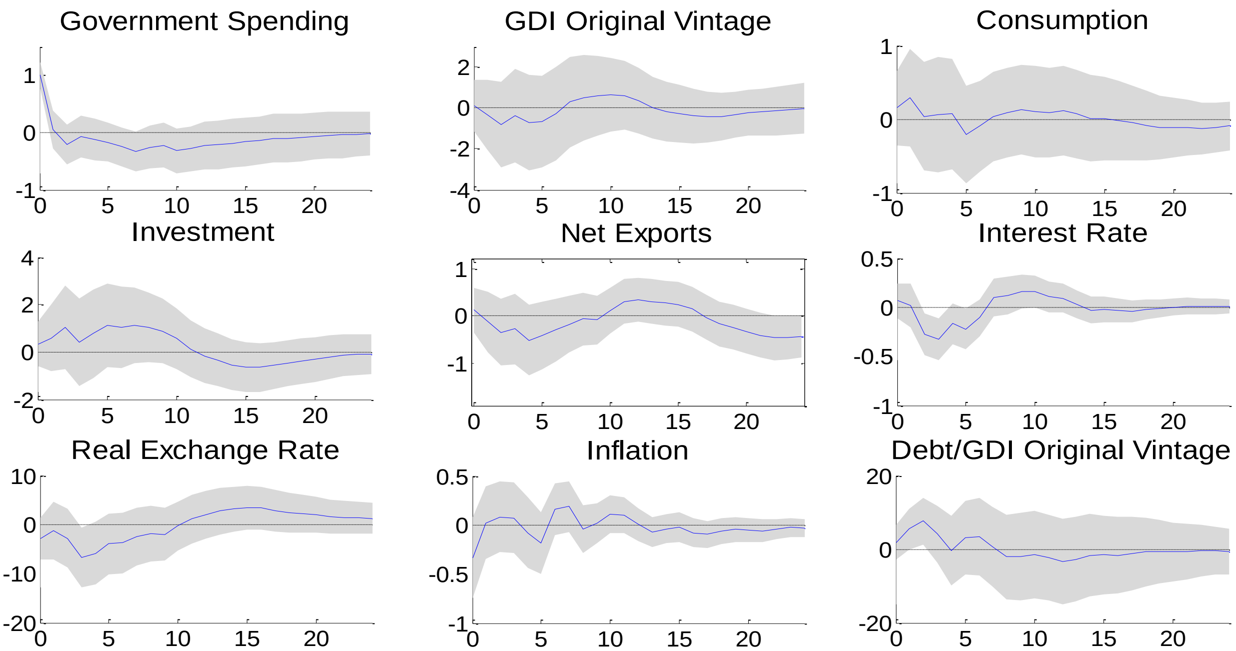
Impulse responses from a vector autoregression identified with the Ramey (2011) method. Solid blue lines indicate the point estimate. Grey area indicates the 90% confidence interval. Horizontal axis indicates quarters. Vertical axes denotes deviations from trend in percent points of trend output (in the case of quantities); percentage deviations from the preshock level (real exchange rate); and deviations from the preshock level in terms of quarterly percentage points (real interest rate and inflation).

Figure 8: Corsetti, Meier, and Müller (2012) Figure 1 With Original-Vintage GDI



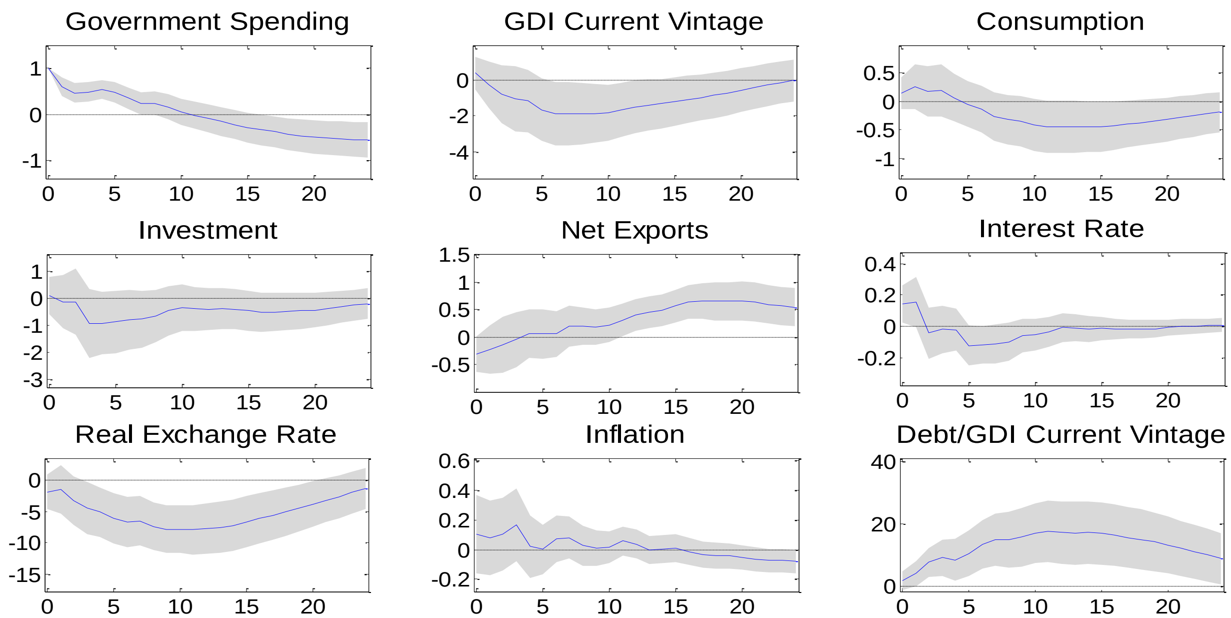
Original vintage is March 2010. Impulse responses from a vector autoregression identified with the Blanchard and Perotti (2002) method. Solid blue lines indicate the point estimate. Grey area indicates the 90% confidence interval. Horizontal axis indicates quarters. Vertical axes denotes deviations from trend in percent points of trend output (in the case of quantities); percentage deviations from the preshock level (real exchange rate); and deviations from the preshock level in terms of quarterly percentage points (real interest rate and inflation).

Figure 9: Corsetti, Meier, and Müller (2012) Figure 2 With Original-Vintage GDI



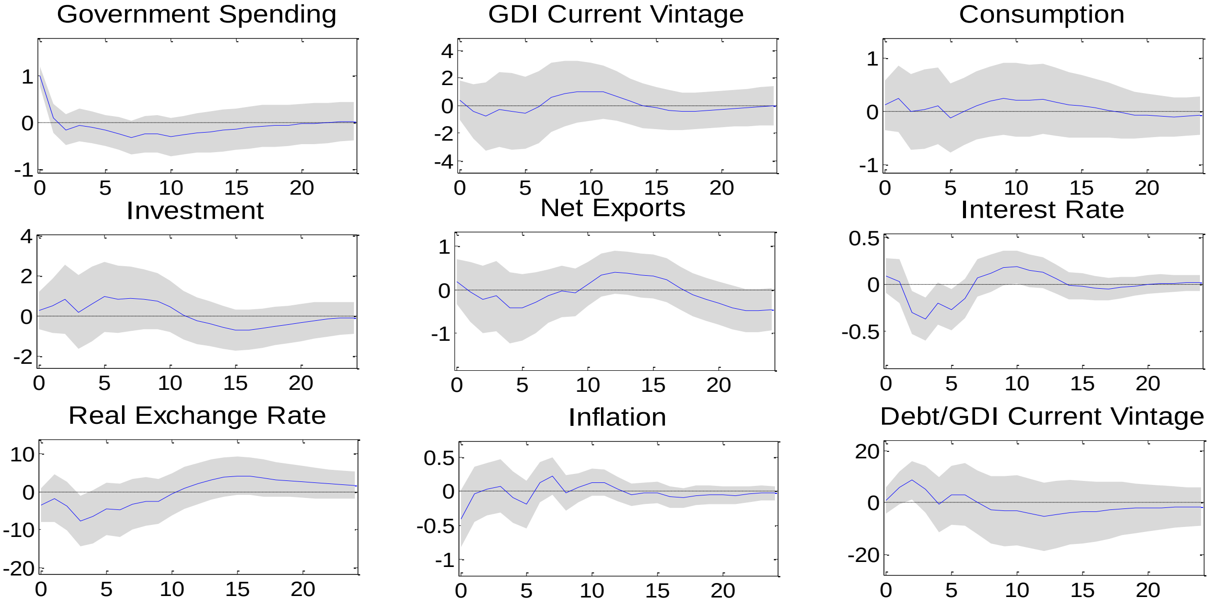
Original vintage is March 2010. Impulse responses from a vector autoregression identified with the Ramey (2011) method. Solid blue lines indicate the point estimate. Grey area indicates the 90% confidence interval. Horizontal axis indicates quarters. Vertical axes denotes deviations from trend in percent points of trend output (in the case of quantities); percentage deviations from the preshock level (real exchange rate); and deviations from the preshock level in terms of quarterly percentage points (real interest rate and inflation).

Figure 10: Corsetti, Meier, and Müller (2012) Figure 1 With Current-Vintage GDI



Impulse responses from a vector autoregression identified with the Blanchard and Perotti (2002) method. Solid blue lines indicate the point estimate. Grey area indicates the 90% confidence interval. Horizontal axis indicates quarters. Vertical axes denotes deviations from trend in percent points of trend output (in the case of quantities); percentage deviations from the preshock level (real exchange rate); and deviations from the preshock level in terms of quarterly percentage points (real interest rate and inflation).

Figure 11: Corsetti, Meier, and Müller (2012) Figure 2 With Current-Vintage GDI



Impulse responses from a vector autoregression identified with the Ramey (2011) method. Solid blue lines indicate the point estimate. Grey area indicates the 90% confidence interval. Horizontal axis indicates quarters. Vertical axes denotes deviations from trend in percent points of trend output (in the case of quantities); percentage deviations from the preshock level (real exchange rate); and deviations from the preshock level in terms of quarterly percentage points (real interest rate and inflation).

Figure 12: Morley and Piger (2012) Figure 3 Replication

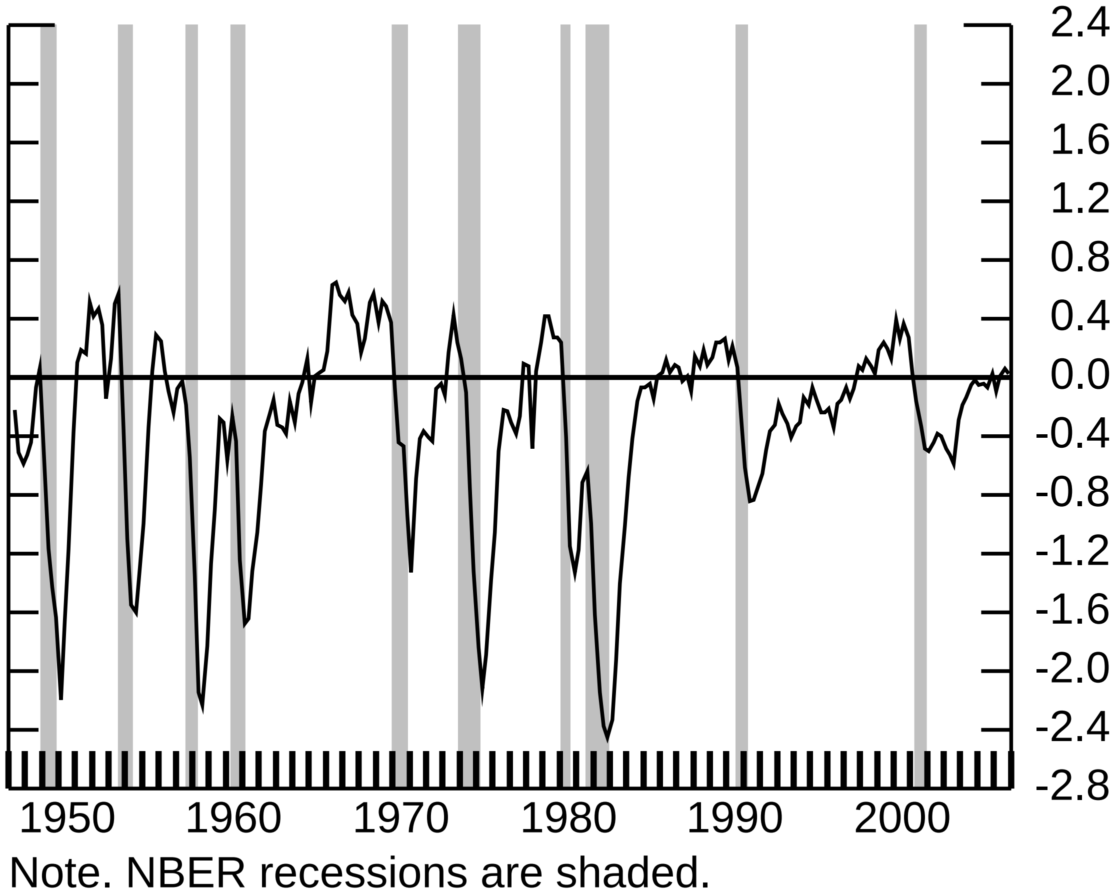
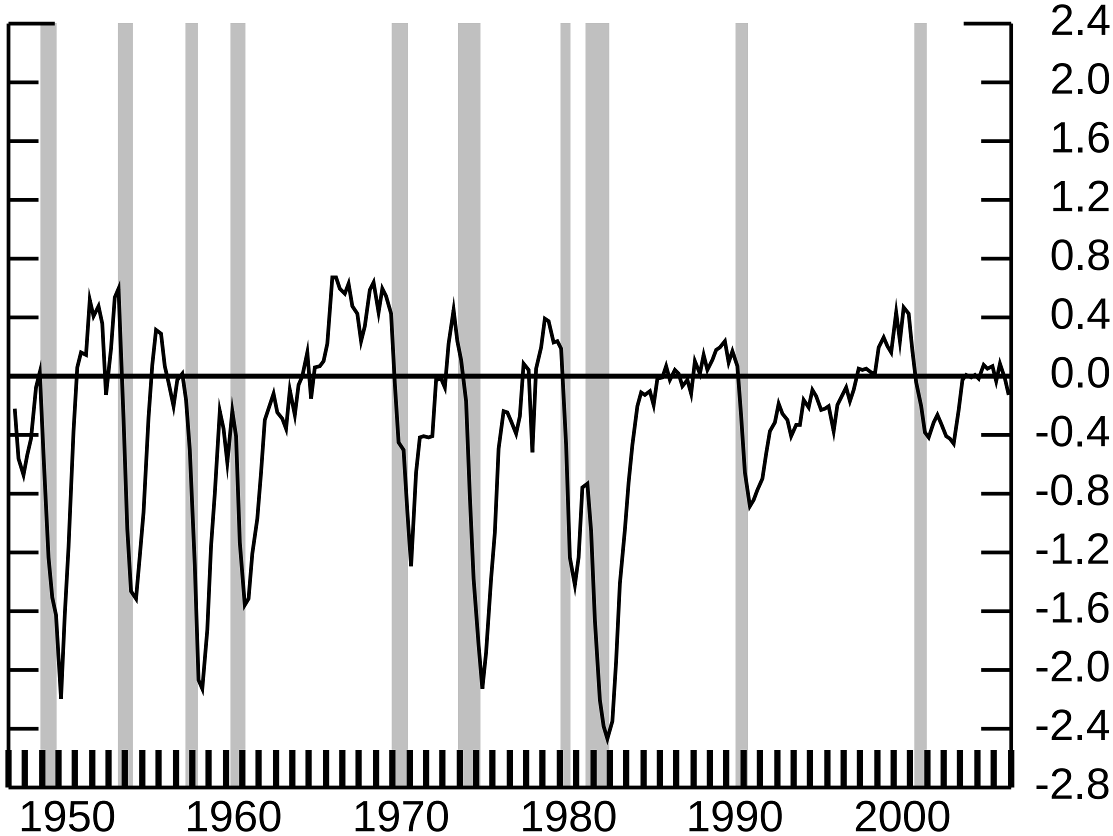


Figure plots the Morley and Piger (2012) model-averaged measure of the business cycle, which is constructed using Bayesian Model Averaging over 33 univariate models. Source: Chang and Li (2015a, Forthcoming).

Figure 13: Morley and Piger (2012) Figure 3 With Current-Vintage GDP



Note. NBER recessions are shaded.

Figure plots the Morley and Piger (2012) model-averaged measure of the business cycle, which is constructed using Bayesian Model Averaging over 33 univariate models.

Figure 14: Morley and Piger (2012) Figure 3 With Original-Vintage GDI

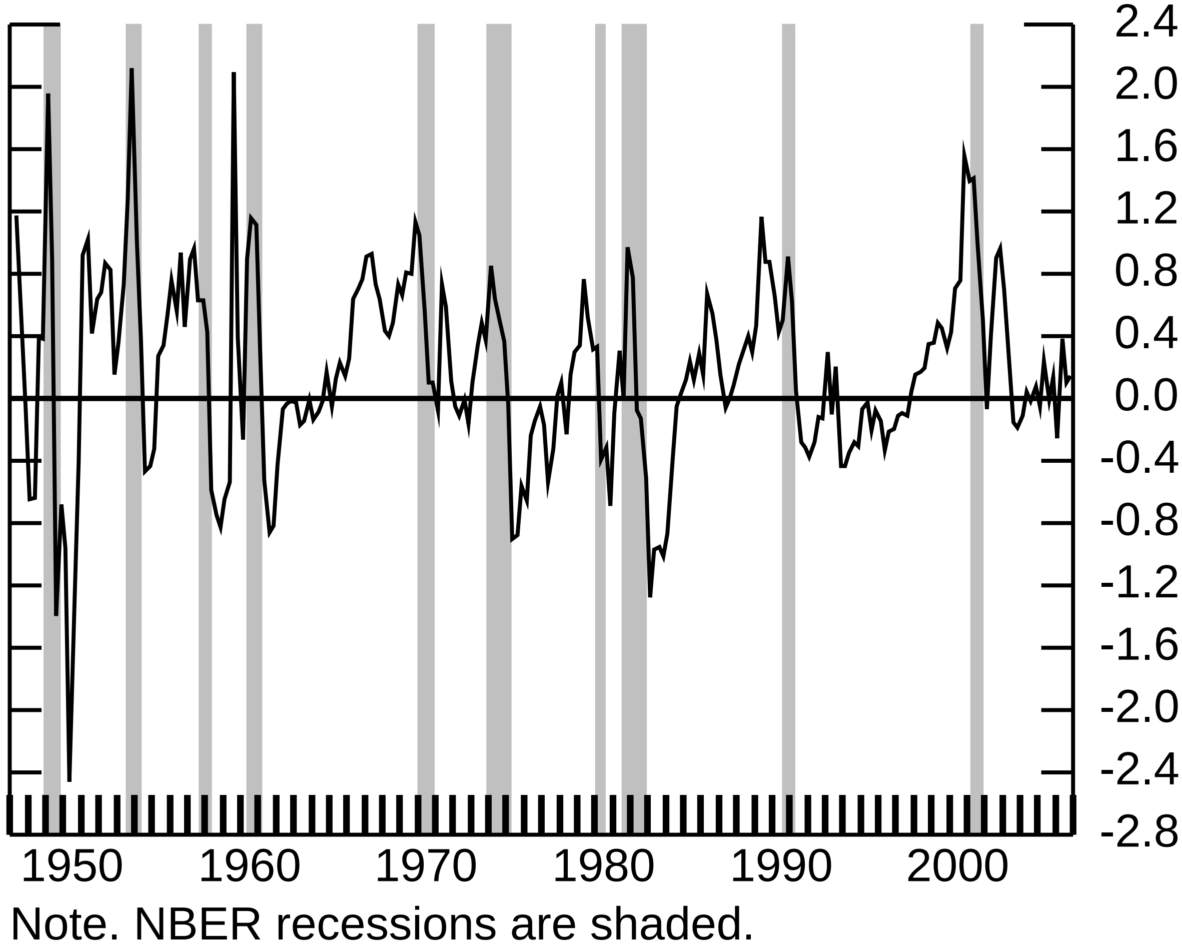
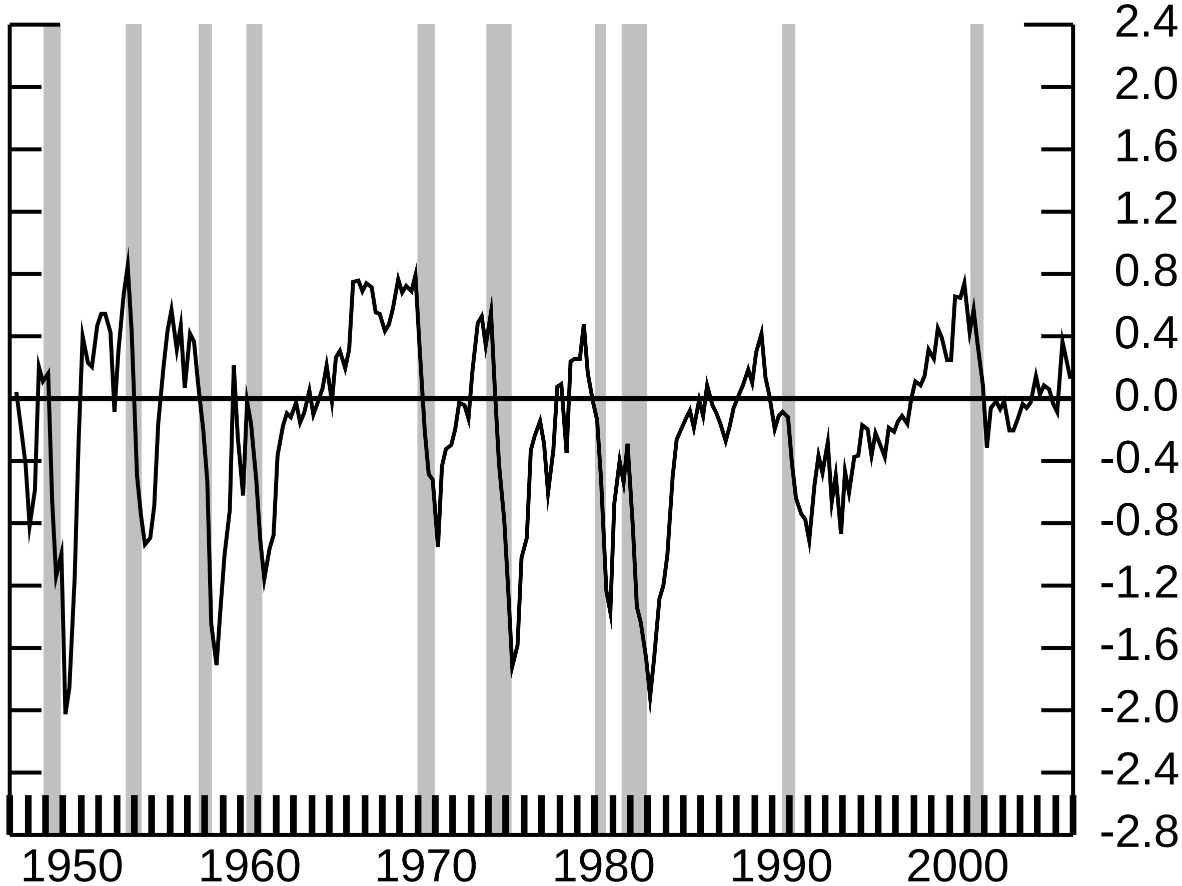


Figure plots the Morley and Piger (2012) model-averaged measure of the business cycle, which is constructed using Bayesian Model Averaging over 33 univariate models.

Figure 15: Morley and Piger (2012) Figure 3 With Current-Vintage GDI



Note. NBER recessions are shaded.

Figure plots the Morley and Piger (2012) model-averaged measure of the business cycle, which is constructed using Bayesian Model Averaging over 33 univariate models.

References

Andrews, Donald W.K., “Tests for Parameter Instability and Structural Change with Unknown Change Point,” Econometrica 61:4 (1993), 821-856.

Auerbach, Alan J., and Yuriy Gorodnichenko, “Measuring the Output Responses to Fiscal

Policy," American Economic Journal: Economic Policy 4:2 (2012), 1-27.

Auerbach, Alan J., and Yuriy Gorodnichenko, "Corrigendum: Measuring the Output Responses to Fiscal Policy," American Economic Journal: Economic Policy 5:3 (2013), 320-322.

Aruoba, Borağan S., Francis X. Diebold, Jeremy Nalewaik, Frank Schorfheide, and Dongho Song, "Improving GDP Measurement: A Measurement-Error Perspective," Journal of Econometrics 191:2 (2016), 384-397.

Barro, Robert J., and Charles J. Redlick, "Macroeconomic Effects from Government Purchases and Taxes," Quarterly Journal of Economics 126:1 (2011), 51-102.

Baumeister, Christiane, and Gert Peersman, "Time-Varying Effects of Oil Supply Shocks on the US Economy," American Economic Journal: Macroeconomics 5:4 (2013), 1-28.

Blanchard, Olivier, and Roberto Perotti, "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output," Quarterly Journal of Economics 117:4 (2002), 1329-1368.

Blanchard, Olivier Jean, and Danny Quah, "The Dynamic Effects of Aggregate Demand and Supply Disturbances," American Economic Review 79:4 (1989), 655-673.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," American Economic Journal: Applied Economics 8:1 (2016), 1-32.

Canova, Fabio, and Luca Gambetti, "Do Expectations Matter? The Great Moderation Revisited," American Economic Journal: Macroeconomics 2:3 (2010), 183-205.

Carey, Colleen, and Stephen H. Shore, "From the Peaks to the Valleys: Cross-State Evidence on Income Volatility Over the Business Cycle," Review of Economics and Statistics 95:2 (2013), 549-562.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, “Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan,” Quarterly Journal of Economics 127:4 (2012), 1755-1812.

Castro, Francisco, Javier J. Pérez, and Marta Rodríguez-Vives, “Fiscal Data Revisions in Europe,” Journal of Money, Credit and Banking 45:6 (2013), 1187-1209.

Chang, Andrew C., and Tyler J. Hanson, “The Accuracy of Forecasts Prepared for the Federal Open Market Committee,” Journal of Economics and Business 83 (2016), 23-43. Previous version as Board of Governors of the Federal Reserve System FEDS Working Paper 2015-062. http://dx.doi.org/10.17016/FEDS.2015.062

Chang, Andrew C., and Phillip Li, “Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say “Usually Not”,” Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System. (2015a). http://dx.doi.org/10.17016/FEDS.2015.083

Chang, Andrew C., and Phillip Li, “Measurement Error in Macroeconomic Data and Economics Research: Data Revisions, Gross Domestic Product, and Gross Domestic Income,” Finance and Economics Discussion Series 2015-102. Washington: Board of Governors of the Federal Reserve System. (2015b). http://dx.doi.org/10.17016/FEDS.2015.102

Chang, Andrew C., and Phillip Li, “A Preanalysis Plan to Replicate Sixty Economics Papers that Worked Half of the Time,” American Economic Review 107:5 (2017), 60-64.

Chang, Andrew C., and Phillip Li, “Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say “Often Not”,” Critical Finance Review (Forthcoming).

Chen, Han, Vasco Curdia, and Andrea Ferrero, “The Macroeconomic Effects of Large-scale Asset Purchase Programmes,” Economic Journal 122:564 (2012), F289-F315.

Clark, Todd E., and Michael W. McCracken, “Averaging Forecasts from VARs With Uncertain Instabilities,” Journal of Applied Econometrics 25:1 (2010), 5-29.

Corsetti, Giancarlo, André Meier, and Gernot J. Müller, “Fiscal Stimulus with Spending Reversals,” Review of Economics and Statistics 94:4 (2012), 878-895.

Croushore, Dean, “Frontiers of Real-Time Data Analysis,” Journal of Economic Literature 49:1 (2011), 72-100.

Croushore, Dean, and Tom Stark, “Is Macroeconomic Research Robust to Alternative Data Sets?” Federal Reserve Bank of Philadelphia Working Paper No. 02-3 (2002).

Croushore, Dean, and Tom Stark, “A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?” Review of Economics and Statistics 85:3 (2003), 605-617.

D’Agostino, Antonello, and Palo Surico, “A Century of Inflation Forecasts,” Review of Economics and Statistics 94:4 (2012), 1097-1106.

Den Haan, Wouter J., and Vincent Sterk, “The Myth of Financial Innovation and the Great Moderation,” Economic Journal 121:553 (2011), 707-739.

Diebold, Francis X., “Comment on The Income- and Expenditure-Side Estimates of U.S. Output Growth,” Brookings Papers on Economic Activity (2010), 107-112.

Favero, Carlo, and Giavazzi, Francesco, “Measuring Tax Multipliers: The Narrative Method in Fiscal VARs,” American Economic Journal: Economic Policy, 4:2 (2012), 69-94.

Faust, Jon, John H. Rogers, Shing-Yi. B. Wang, and Jonathan H. Wright, “The High-Frequency Response of Exchange Rates and Interest Rates to Macroeconomic Announcements,” Journal of Monetary Economics 54:4 (2007), 1051-1068.

Faust, Jon, John H. Rogers, and Jonathan H. Wright, “Exchange Rate Forecasting: The Errors We’ve Really Made,” Journal of International Economics 60:1 (2003), 35-59.

Feng, Shuaizhang, and Yingyao Hu, "Misclassification Errors and the Underestimation of the US Unemployment Rate," American Economic Review 103:2 (2013), 1054-1070.

Fixler, Dennis J., and Bruce T. Grimm, "The Reliability of the GDP and GDI Estimates," Survey of Current Business 88:2 (2008), 16-32.

Fleischman, Charles A., and John M. Roberts, "From Many Series, One Cycle: Improved Estimates of the Business Cycle from a Multivariate Unobserved Components Model," Board of Governors of the Federal Reserve System FEDS Working Paper 2011-46 (2011).

Gabaix, Xavier, "The Granular Origins of Aggregate Fluctuations," Econometrica 79:3 (2011), 733-772.

Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," Journal of Political Economy 86:6 (1978), 971-987.

Hamermesh, Daniel S., "Viewpoint: Replication in Economics," Canadian Journal of Economics 40:3 (2007), 715-733.

Hansen, Peter R., Asger Lunde, and James M. Nason, "The Model Confidence Set," Econometrica 79:2 (2011), 453-497.

Heutel, Garth, "How Should Environmental Policy Respond to Business Cycles? Optimal Policy Under Persistent Productivity Shocks," Review of Economic Dynamics 15:2 (2012), 244-264.

Hodrick, Robert J., and Edward C. Prescott, "Postwar U.S. Business Cycles: An Empirical Investigation," Journal of Money, Credit and Banking 29:1 (1997), 1-16.

Ireland, Peter N., "On the Welfare Cost of Inflation and the Recent Behavior of Money Demand," American Economic Review 99:3 (2009), 1040-1052.

Inoue, Atsushi, and Barbara Rossi, "Identifying the Sources of Instabilities in Macroeconomic Fluctuations," Review of Economics and Statistics 93:4 (2011), 1186-1204.

Kilian, Lutz, “Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market,” American Economic Review 99:3 (2009), 1053-1069.

Koenig, Evan F., Sheila Dolmas, and Jeremy Piger, “The Use and Abuse of Real-Time Data in Economic Forecasting,” Review of Economics and Statistics 85:3 (2003), 618-628.

Kormilitsina, Anna, “Oil Price Shocks and the Optimality of Monetary Policy,” Review of Economic Dynamics 14:1 (2011), 199-223.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen, “The Aggregate Demand for Treasury Debt,” Journal of Political Economy 120:2 (2012), 233-267.

Kydland, Finn E., and Edward C. Prescott, “Business Cycles: Real Facts and a Monetary Myth,” Federal Reserve Bank of Minneapolis Quarterly Review (1990), 3-18.

Landefeld, J. Steven, “Comment on The Income- and Expenditure-Side Estimates of U.S. Output Growth,” Brookings Papers on Economic Activity (2010), 112-123.

Landefeld, J. Steven, Eugene P. Seskin, and Barbara M. Fraumeni, “Taking the Pulse of the Economy: Measuring GDP,” Journal of Economic Perspectives 22:2 (2008), 193-216.

Mankiw, N. Gregory, and Matthew D. Shapiro, “News or Noise? An Analysis of GNP Revisions,” Survey of Current Business 66 (1986), 20-25.

Mark, Nelson C., “Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability,” American Economic Review 85:1 (1995), 201-218.

Mavroeidis, Sophocles, “Monetary Policy Rules and Macroeconomic Stability: Some New Evidence,” American Economic Review 100:1 (2010), 491-503.

Mertens, Karel, and Morten O. Ravn, “Understanding the Aggregate Effects of Anticipated and Unanticipated Tax Policy Shocks,” Review of Economic Dynamics 14:1 (2011), 27-54.

Mertens, Karel, and Moten O. Ravn, “The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States,” American Economic Review 103:4 (2013), 1212-1247.

Michalski, Tomasz, and Gilles Stoltz, “Do Countries Falsify Economic Data Strategically? Some Evidence that They Might,” Review of Economics and Statistics 95:2 (2013), 591-616.

Morley, James, and Jeremy Piger, “The Asymmetric Business Cycle,” Review of Economics and Statistics 94:1 (2012), 208-221.

Nakov, Anton, and Andrea Pescatori, “Oil and the Great Moderation,” Economic Journal 120:543 (2010), 131-156.

Nalewaik, Jeremy J., “The Income- and Expenditure-Side Estimates of U.S. Output Growth,” Brookings Papers on Economic Activity (2010), 71-106.

Nalewaik, Jeremy J., “Estimating Probabilities of Recession in Real Time Using GDP and GDI,” Journal of Money, Credit and Banking 44:1 (2012), 235-253. Also as Board of Governors of the Federal Reserve System FEDS Working Paper 2007-07.

Nalewaik, Jeremy J., “Missing Variation in the Great Moderation: Lack of Signal Error and OLS Regression,” Working Paper (2014). Also as Board of Governors of the Federal Reserve System FEDS Working Paper 2008-15.

Neumark, David, “The Employment Effects of Recent Minimum Wage Increases: Evidence from a Pre-Specified Research Design,” NBER Working Paper No. 7171 (1999).

Neumark, David, “The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design,” Industrial Relations: A Journal of Economy and Society 40:1 (2001), 121-144.

Noorbakhsh, Farhad, "International Convergence or Higher Inequality in Human Development?" UNU-Wider Research Paper 2006/15 (2006).

Orphanides, Athanasios, "Monetary Policy Rules Based on Real-Time Data," American Economic Review 91:4 (2001), 964-985.

Orphanides, Athanasios, and Simon van Norden, "The Unreliability of Output-Gap Estimates in Real Time," Review of Economics and Statistics 84:4 (2002), 569-583.

Phillips, P. C. B., and S. Ouliaris, "Asymptotic Properties of Residual Based Tests for Cointegration," Econometrica 58:1 (1990), 165-193.

Ponomareva, Natalia, and Hajime Katayama, "Does the Version of the Penn World Tables Matter? An Analysis of the Relationship Between Growth and Volatility," Canadian Journal of Economics 41:1 (2010), 152-179.

Rangel, José Gonzalo, "Macroeconomic News, Announcements, and Stock Market Jump Intensity Dynamics," Journal of Banking & Finance 35:5 (2011), 1263-1276.

Ramey, Garey, and Valerie A. Ramey, "Cross-Country Evidence on the Link Between Volatility and Growth," American Economic Review 95 (1995), 1138-1151.

Ramey, Valerie A., "Identifying Government Spending Shocks: It's all in the Timing," Quarterly Journal of Economics 126:1 (2011), 1-50.

Reis, Ricardo, and Mark W. Watson, "Relative Goods' Prices, Pure Inflation, and The Phillips Correlation," American Economic Journal: Macroeconomics 2:3 (2010), 128-157.

Romer, Christina D., and David H. Romer, "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks," American Economic Review 100:3 (2010), 763-801.

Rosenthal, Robert, "The File Drawer Problem and Tolerance for Null Results," Psychological Bulletin 86:3 (1979), 638-641.

Schmitt-Grohé, Stephanie, and Martín Uribe, "Business Cycles with a Common Trend in Neutral and Investment-Specific Productivity," Review of Economic Dynamics 14:1 (2011), 122-135.

Schmitt-Grohé, Stephanie, and Martín Uribe, "What's News in Business Cycles," Econometrica 80:6 (2012), 2733-2764.

Taylor, John B., "Discretion Versus Policy Rules in Practice," Carnegie-Rochester Conference Series on Public Policy 39 (1993), 195-214.

Whelan, Karl, "A Guide To U.S. Chain Aggregated NIPA Data," Review of Income and Wealth 48:2 (2002), 217-233.

Whittaker, E.T., "On a New Method of Graduation," Proceedings of the Edinburgh Mathematical Society 41 (1923), 63-75.

Wolff, Hendrik, Howard Chong, and Maximilian Auffhammer, "Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index," Economic Journal 121:533 (2011), 843-870.

Zucman, Gabriel, "The Missing Wealth of Nations: Are Europe and the U.S. Net Debtors or Net Creditors?" Quarterly Journal of Economics 128:3 (2013), 1321-1364.