Quantitative Easing and Economic Growth in Japan: A Meta-Analysis [*]

Alexandra Ferreira-Lopes
ISCTE-IUL and BRU-IUL

Pedro Linhares
ISCTE-IUL

Luís Filipe Martins
ISCTE-IUL, BRU-IUL, and CIMS-University of Surrey

Tiago Neves Sequeira
Univ. Coimbra and CeBER

This is the peer reviewed version of the following article: [Ferreira-Lopes, A., Linhares, P., Martins, L. F., & Sequeira,T. N. (2022). Quantitative easing and economic growth in Japan: A meta-analysis.Journalof Economic Surveys,36:235–268.], which has been published in final form at [https://doi.org/10.1111/joes.12449]. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract

The Bank of Japan is the central bank that has been using unconventional monetary policies (UMP) for the longest time. We present an original meta-probit analysis for the period 2001-2020, using 45 studies, which focuses on the literature that has been studying the effect of those policies on Japanese output growth. We use impulse response functions in Vector Auto Regression (VAR) type models to assess if the effects on output of UMP are significant (or not), and if significant, positive or negative. Funnel asymmetry and precision effect tests do not provide convincing evidence of publication bias. Additionally, we also do not find a consensus regarding the output growth effects during the quantitative easing years. A thorough meta-probit analysis suggests that certain variables included in the sample such as industrial production (as a proxy for output) and the price level – as well as different specifications in the data used – have a greater effect on the probability of reporting statistically significant (positive) effects of quantitative easing on output growth. Additionally, other variables, like if it is an older study, if the study includes several quantitative easing (QE) programs, if it is published in a journal, if it is using monthly or daily

[*] Alexandra Ferreira-Lopes, Luís Filipe Martins, and Tiago Neves Sequeira would like to thank the Portuguese National Science Foundation (FCT) and FEDER/COMPETE for funding, through projects UIDB/00315/2020 and UIDB/05037/2020. The authors would also like to thank participants at the 25th Annual Conference of the Computing in Economics and Finance Society and the 50th Annual Conference of the Money, Macro, and Finance Society, seven referees, the associate editor, the managing editor, as well as Joaquim Ramalho for very useful suggestions. The usual disclaimer applies.

data, if it is using one standard deviation in the confidence intervals, if the study does not use a monetary supply or a money base variable, but if the shock is applied to the monetary base, also increase the probability of reporting statistically significant (positive) effects of quantitative easing on output growth. However, one of the most important (policy) implications of our study is that until now the evidence for a significant real effect of unconventional monetary policy is weak, since concerning true effects, we obtain mixed results.

Keywords: Quantitative easing, Bank of Japan, monetary policy, economic growth, meta-probit.
JEL Codes: E52, E58; O11; O42; O53.

1. Introduction

The Bank of Japan has been employing measures of unconventional monetary policy (UMP) for over 20 years, and is by far the central bank (e.g., compared with the United Kingdom, the USA, or the Euro area) that has been applying these policies for longer.[2] These measures, which involve mostly large-scale purchases of long-run financial assets, were set in place especially to stimulate the sluggish Japanese economy. However, the Japanese case is still perceived by the majority of the audience – policy makers, researchers, investors, and even the public in general – as one of the most notorious histories of an economy incapable of detaching itself from stagnation, regardless of the efforts to do so. In fact, according to data from the Penn World Tables (PWT), version 10.0, the Japanese economy presented an average economic growth rate of only 0.2% between 2000 and 2019 (Feenstra *et al.*, 2015).

A substantial bulk of empirical work has appeared on the effects of the measures of unconventional monetary policy on several fundamental variables for the Japanese economy. Controversy on the effects of the unconventional monetary policy followed by the Bank of Japan has become common. For example, Montgomery and Volz (2019) found that UMP policies were effective in establishing a bank's lending channel. On the other hand, Lombardi *et al.* (2018) argued that although UMP can prevent economic collapse, it is not designed to promote long-run economic growth.

We study the UMP of the Bank of Japan (BoJ) and its impact on GDP and/or economic growth. Specifically, we rely on a meta-analysis of 45 studies covering a publishing period between 2006 and 2020 to draw valid and statistically valuable conclusions. We examine a selection of empirical literature, the characteristics of that literature, as well as methods, time periods, and variables that assess and measure the impact of the policies undertaken by the Bank of Japan on GDP and/or its growth. We use meta-analysis for our work, since the bulk of empirical studies that have computed the impact of the BoJ's monetary policy on GDP and/or economic growth, namely quantitative easing, is substantial. Some recent non-quantitative surveys of the literature on UMP exist (Dell'Ariccia *et al.*, 2018; Fatouh, 2016; Papadamou *et al.*, 2020; Poter and Smets, 2019; and Turner *et al.*, 2019), some of which includes countries other than Japan. Our work clearly differs from this set of literature, as we provide a meta-regression analysis, thereby providing quantitative evidence that may shed light on the controversy of the real effects of the UMP of the Bank of Japan.[3] Our database also speaks to this controversy, as we have 47.3% of the studies

[2] This has been noted also by Bhattarai and Neely (2018) in a non-quantitative revision of the literature.

[3] Some authors have also been concerned with cross-national effects of the UMP (Gros *et al.*, 2015; Gambacorta *et al.*, 2014; Buch *et al.*, 2019 and Ganelly and Tawk, 2019). Overall, results tend to indicate some reasons that may determine small real effects of UMP, either because spillovers to other countries are small or because cross-border effects may have effects opposite to the national ones.

reporting positive (significant) effects, 8.2% reporting negative (significant) effects, and 44.5% reporting non-significant effects of QE on output.

As will become clear from the review of the literature on meta-analysis of monetary policy and in particular on the policy applied in Japan, our contribution is twofold. Firstly, we contribute to the meta-analysis of monetary policies, specifically UMP such as quantitative easing, which is clearly still an overlooked issue in the literature. To our knowledge there is only one very recent meta-analysis of the real effects of UMP (including Japan), that of Papadamou *et al.* (2019). Their paper, however, examines only 16 primary studies and includes countries other than Japan. The authors used a more traditional econometric approach than ours (ordinary least squares (OLS) and weighted least squares (WLS)). Our meta-analysis method – meta-probit – has never been used before in this type of literature, which has important advantages in this context. With a probit-type model we capture the overall effect of the monetary policy shock on output by assigning the value of "1" if the effect size is statistically significant and positive and "0" otherwise. Specifically, it expands the number of estimated effect-sizes in Vector Auto Regression (VAR)-type models. In this context, the VAR models are the most appropriate class of specifications to use, not only because they take all the variables in the system as endogenous, and provide estimates of future economic impacts due to the occurrence of the structural shocks of interest, but also because they are the most used in primary studies of the effect of monetary policies in the case of Japan.

In addition to the probit-meta approach, we implement new developments in meta-analysis, namely the cluster bootstrap t-tests and the model averaging smoothing estimation schemes. These prominent recent meta-analyses in economics allow us to draw more robust conclusions. The bootstrap t-test confirms the standard and the cluster robust t-tests and the averaging of models' point estimates summarizes the information contained in the individual estimated models.

Secondly, we contribute by presenting a systematic synthesis on the effect of the BoJ's quantitative easing on the Japanese real economy based on the literature that has been analysing these effects. That is, our study uses a completely new meta-sample. In this sense, our results have important policy implications, not only for the Japanese monetary authority, but also for other countries' monetary authorities, with the necessary adaptations. This is because quantitative easing has been spreading as a policy response to the challenges arising from the sub-prime crises in 2008/09 and the subsequent Eurozone debt crisis.

Our results provide no evidence of publication bias, nor of a consensus over the output growth effects during the quantitative easing years, since the average effect is sometimes non-significant and different specifications lead to diverse quantitative true effects. These results suggest that QE policies may have no effect on average on real output variables, in both the very short and medium runs, and countries may not rely on those policies to relaunch the economy when in crisis.

Additionally, the meta-probit analysis suggests that industrial production (used as a proxy for output instead of other proxies) and the price level (present in the estimation as opposed to not using it), affect more the probability of reporting statistically significant (positive) effects of quantitative easing on output growth. Other specifications used, such as being an older study, if the study includes several QE programs (instead of just the 1st QE), if it is published in a journal (instead of a working paper only), if it is using monthly or daily data (instead of quarterly data), if it is using one standard deviation in the confidence intervals; also affect more the probability of reporting statistically significant (positive) effects of quantitative easing on output growth.

Additionally, if in a study either a monetary supply variable or a money base variable is chosen over other types of monetary variables, the probability that a study finds a positive and significant effect size falls dramatically. However, if the shock used to produce the output estimations is of a money base type (as opposed to other monetary policy instruments), then the probability of reporting statistically significant (positive) effects of quantitative easing on output growth increases. One can then refer to the overall average marginal effect of choosing money base over other types of monetary variable as the difference of those two opposing effects, with the net difference yielding a negative result. Because the monetary variable is a crucial one to produce the output estimations, it is not unreasonable to state that the type of monetary variable used affects the behaviour of the impulse response functions when a monetary policy shock exists, thus affecting the output estimation in terms of its sign and significance. This has direct policy implications for the variable chosen as a monetary policy instrument.

On the contrary, variables like the studies' impact ranking, number of observations, whether an author is associated with the Bank of Japan, the type of VAR-type model used, whether the output variable is measured in levels or in first differences, and the inclusion of an interest rate variable, an exchange rate variable, or a bond variable in the model, do not seem relevant to the sign and significance of the effect size of QE on output for the Japanese economy.

One of the most important (policy) implications of our study is that until now the evidence for a significant real effect of unconventional monetary policy is weak, since with regard to true effects, we obtain mixed results.

This study has the following structure: Section 2 establishes the context of the study with regard to the Japanese case and presents the existing cases of meta-analysis literature focused on monetary policy transmission. Section 3 gives an account on how data were collected from the literature on Japanese monetary policy, and how they were treated and organized. Section 4 uses the dataset to conduct a publication bias screening based on funnel plots, precision effect tests, and linear regressions in order to screen for biased results in the selected published literature. Section 5 presents a series of probit estimations that seek to reveal whether the choice or presence of certain elements that characterize that same literature, regarding the type of data used, or other methodological aspects, are able to predict what they report. Section 6 concludes.

2. The Context of the Study: Reviewing the Japanese Case and the Literature on Meta-analysis

2.1. The Evolution of the Japanese Economy and the Need for Unconventional Monetary Policies

The last 20 years of the Japanese economy can be described as a series of successive attempts to recover from what is generally designated as *Japan's lost decade*, the 1990s. This decade is considered a significant turning point for the Japanese economy, characterized by a long-lasting recession, which ended in the early 2000s in a combination of negative output gap (as well as sluggish economic growth) and moderate deflation. The efforts to reverse this scenario has persisted throughout the recent economic history of Japan, and are considered as one of the greatest challenges for national policy makers – with special responsibilities for the institution that runs Japan's monetary policy, the BoJ.

In order to reverse the scenario of the 1990s, the BoJ engaged in what is known to date to be an unconventional type of policy framework, replacing the main policy tools and the policy targets. The unconventional monetary policy approach during this period is known as quantitative easing (QE). Since 2001 there have been three programs that fall under the category of QE, with the last of these implemented (and still in place) designated as the qualitative and quantitative easing (QQE) program. Using the definition of Ugai (2015), a QE program can be regarded as the work of two mechanisms that operate through both sides of a central bank's balance sheet. According to the author, if a central bank operates through the purchase of risky assets in order to diminish the imbalances of the financial markets, then the central bank is using the asset side of its balance sheet; if the central bank engages in large-scale operations of government bond purchasing, which will force the monetary base of the economy to expand in a first stage, then the liability side of the same balance sheet is being used.

2.2 Monetary Policy Transmission under Meta-analysis

Regarding the literature that could serve as a methodological object of comparison, we found a small number of studies that employ meta-analysis focusing on monetary policy transmission and that, more specifically, simultaneously distinguish between types of VAR methodologies employed and make use of an effect size based on impulse response functions[4] (IRFs). However, none of these studies uses meta-probit to analyse the impact of monetary policy in order to find the main determinants that imply a larger probability of studies reporting significant effects on growth and inflation. Table 1 summarizes some of the literature addressing monetary policy transmission. For instance, Grauwe and Storti (2004) used a meta-regression to infer the factors that could justify the variation of results reported in the literature regarding the impacts of monetary policy shocks in the output and the price level. They find large variation in the results reported concerning the estimations for output, stating as well that part of that variation could be explained by whether the authors used VAR or SVAR techniques. In a similar fashion, Ridhwan *et al.* (2010) conducted a meta-analysis that revealed an accentuated variance of estimated output effects taken from the literature, regarding both the speed and magnitude of transmission.

Uusküla and Pitzel (2007) found some evidence for countries within the EU-15 that greater financial depth is positively correlated with a stronger transmission of monetary shocks. These authors took three alternative variables that measure financial depth for each country and assessed their correlations with the corresponding monetary shock impacts on output and prices based on the IRFs reported in the literature. Rusnak *et al.* (2013) employed a mixed-effects multilevel model that sought to capture the reasons behind the price puzzle patterns found across the literature. The findings in their study suggest that patterns observed in the empirical estimates often do not align with what the theory postulates. The authors also found that having more sample observations weakens the response of the price level and that the reported estimations do vary depending on the VAR specification and output proxy used.

Instead of employing regressions in their research, Havranek and Rusnak (2013) opted for a Bayesian model averaging (BMA) method in order to draw conclusions on the speed of monetary transmission. The authors found that a "*best-practice*" model based on the results reported by their BMA approach shortens the average time of shock transmission in the price level considerably when compared with the average taken from the results in the literature. Moreover, these authors also found that data and methodology play a role in explaining the variation of results in the literature. Studies that use monthly data and report strictly decreasing impulse responses are prone to produce evidence of a slower transmission, whereas studies that report *hump-shaped* impulse responses tend to report a faster transmission of monetary policy shocks into the price level.

In a meta-study with 16 published articles, Papamadou *et al.* (2019) uses OLS and WLS meta-regressions to study the effects of unconventional monetary policy on output and inflation. Their

[4] The studies cited also used meta-analysis in order to unveil cross-country heterogeneity of the results. The variables included are based on economic features such as openness to foreign trade or proxies for financial development.

study includes Japan among other countries and analyses the empirical studies that have used VAR techniques. Their findings indicate that studies that used factor-augmented VAR (FAVAR) techniques exhibit a higher response of output to unconventional monetary policy shocks in all time horizons. Studies about European countries lower the output responses. This is the closest study to our own, although it focus on several countries and does not use meta-probit.

Fabo *et al.* (2020) use 54 studies for the US, UK, and Euro Area to analyse the effect of UMP on output and/or inflation. From all the three regions that the authors analyse, the results for the US are better regarding the positive effect of QE on output and also on the price level. The authors also found that 88% of the studies present a significant effect of QE on output, while 84% present a significant effect of QE on prices.

Table 1: Effect Sizes and Countries/Regions used by Meta-analysis Studies on Monetary Policy

Authors	Effect size(s) registered	Countries
Grauwe and Storti (2004)	1% increase of the interest rate in the output and the price level, caught at the 1st and 5th year.	Austria, Belgium, Denmark, *Emerging Countries*, *Eurozone*, Finland, France Germany, Greece, Ireland, Italy, Japan, Luxembourg, Netherlands, Portugal, Spain, Sweden, UK, and US.
Uusküla and Pitzel (2007)	Maximum value attained by the monetary policy shock in output and price levels.	EU-15 countries.
Ridhwan *et al.* (2010)	1% increase of the interest rate in the output caught at the 14th and 16th quarters and at the maximum and minimum levels.	US, *Eurozone,* and European Union (*Non-Eurozone*).
Rusnak *et al.* (2013)	1% increase of the interest rate in the price level caught at the 3rd, 6th 12th, 18th, and 36th months, and at the maximum and minimum levels.	Australia, Brazil, Bulgaria, Canada, Czech Republic, Denmark, Estonia, *Euro Area*, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Korea, Latvia, Lithuania, Malaysia, New Zealand, Philippines, Poland, Romania, Slovakia, Slovenia, Spain, Thailand, Turkey, UK, and

Havranek and Rusnak (2013)	1% increase of the interest rate in the price level caught at the minimum value (right after the local maximum) for *humped-shaped responses* and at the last period available for strictly decreasing responses.	(Same as in Rusnak *et al.* 2013)
Papadamou *et al.* (2019)	1% increase of the output and prices caught at the maximum values and at the 4th, 8th, and 16th quarters.	Canada, Czech Republic, Denmark, Euro Area, Hungary, Japan, Norway, Poland, Romania, Sweden, Switzerland, UK, and US.
Fabo *et al.* (2020)	Using a QE shock to 1% of the country's pre-QE level of GDP, it increases the level of output by 0.24% and of prices by 0.19% at the peak. The average cumulative effect on output is 58% and on prices is 63% of the peak effect.	US, UK, and the Euro Area

3. Empirical Strategy

In this section, we present the construction and a description of our database, as well as some descriptive statistics.[5]

3.1 Methodology – Construction of the Database

Often the simple meta-analysis includes real-valued size effects that come from the studies (in Tables or in the text). For real-valued effects, it is then standard to devise a least squares model specification and estimation. The goal in our paper is to evaluate the Japanese QE monetary policies, and the studies in which we find this analysis performed typically use VAR-type models. Contrary to the standard approaches, in VAR-type models the impact of policies is evaluated through the impulse-response functions and the Granger-causality test. That is, those studies rarely have real-valued point estimates, but instead present IRFs with the confidence sets. This is the reason why we have a binary-type size effect: from the IRFs and the Granger-causality test we can infer whether or not the policy was effective. More details about why and how we

[5] In Sections 3, 4, and 5 we follow the guidelines for meta-analysis in economics of Havránek *et al.* (2020).

construct these kinds of effect sizes and measure precision are presented next. We search for all the studies available in Google Scholar, RePEC, B-on, and Scopus (which has a wider coverage than ISI Web of Science) having the following features: (1) attempt to respond to the question of whether the Japanese monetary policy was effective in promoting economic growth during the QE periods, even if this was not the main question of the study; and (2) make use of VAR or related methodologies, in which their results follow from impulse response functions (IRFs).

As an additional criterion, the impulse response functions must account for a shock caused by a monetary policy tool that conveys its impact onto an output proxy. For example, are the monetary policy shocks affecting the Japanese economy positively or negatively?; what is the magnitude and duration of such impacts?; can the authors identify transmission channels through which those impacts are conveyed?; what can be said about the statistical robustness of these IRFs?

Another criterion was that a study should include in its period of analysis at least one year of quantitative easing. This excludes studies based on samples prior to 2002.

By imposing these features upon every study selected, we account for a certain degree of homogeneity within that pool, thus creating the necessary basis for comparability between studies. The following set of words were entered into the above mentioned search engines: “Japan”, “Economic Growth”, “Output”, “VAR”, “Effect”, “Effectiveness”, “Impact”, “Impulse Response”, “Monetary Policy”, “Quantitative Easing”, “Transmission Channels”, “Transmission Mechanisms”, “Zero Interest Rate Policy”, and “Zero Lower Bound”. The second step was *snowballing* from the previous search results[6].

The data were collected from 45 studies (see the Online Appendix B), registering a total amount of 146 impulse response estimations, and covering a publishing period from 2006 to 2020 (August).

Each entry in the database corresponds to a single set of information, which intends to register fundamental characteristics of an impulse response of the output variable to a disturbance in a given monetary policy variable. As the information reported in impulse response functions is often not quantified and summarized in a systematic manner, we had to collect this information from the graphical representations available. For this a ruler was recurrently used for support, although there are numerous factors that pose hindrances to its good practice (see, e.g., Rusnak *et al.*, 2013). Other exclusions due to measurement problems and potential selection bias were the following: the impulse responses that are presented in robustness sections[7]; estimates that the researcher(s) disregard from the outset as being irrelevant and/or justifying their computation

[6] *Snowballing* is to continuously look at the references found in the relevant studies and reverse *snowballing* it to search in *Google Scholar* and other platforms for the papers that have cited a given study.

[7] These exercises tend to support or validate the researcher’s main conclusions, and including them could create a bias in the results found in the meta-regression.

simply to support a premise; and estimates produced via data simulation, panel VAR, or estimates reported in 3D representations.

3.2 Description of the Database

The content drawn from the study selection is used to build the database that serves our meta-regression analysis. There are four broad groups of information – *Authorial Information*, *Data*, *Methodological Specifications,* and *Estimates* (see Table 2). For each category a brief description is provided.

Table 2: Summary of the Information Collected to Build the Database

Authorial Information	**Data**
- Authors	- QE Programs within the Analysis' Timeframe
- Year of Publication	- Periodicity of the Time Series
- Type of Publication	- Number of Observations of the Analysis' Timeframe
- Association with the BoJ	- Midpoint of the Study's Timeframe
Methodological Specifications	**Estimates**
- Empirical Method	- Signal of the Shock's Impact on Output
- Variable(s) that measure Output	- Persistence of the Shock's Impact on Output
- Other Variables used in the Regression	- Magnitude – Value of the Shock's Impact on Output
- Type of Shock (1) and (2)	
- Confidence Intervals	
- Output Variable in Levels or in First Differences?	
- Monetary Policy Variable in Levels or in First Differences?	
- Is the Shock Employed at a Specific Date?	
- The Date of the Beginning of the Shock (if applicable)	

3.2.1 Authorial Information

Regarding the variables whose designation name is not self-contained, we have:

- *Year of publication* –the publication date of each study.
- *Type of publication* –published articles, working papers, or mimeos. Additionally, published papers are distinguished between general and those specialized in monetary themes.

- *Are the author(s) associated with the Bank of Japan?* – “yes/no” dichotomy.

3.2.2 Data

This group comprehends:

- *QE programs within the analysis’ timeframe* – because of the existence of different timeframes, we group them in the following way: one group comprises the studies that comprehend only the first QE program; the other group comprises studies that do not analyse exclusively the first QE program[8], as expected, most of the studies analyse the first QE program alone.
- *Periodicity of the time series* –daily, monthly, and/or quarterly.
- *Number of observations of the analysis’ timeframe* – we labelled it in the following way: lower than 50; between 50 and 100; and greater than 100.

3.2.3 Methodological Specification

- *Empirical method* – we observed a very extensive array of variations to the VAR methodology, including Bayesian inference methods (see Table 3). *Switching* models present an internal mechanism that allows distinguishing between the zero interest rate period (ZIRP) and normal regimes. Due to the scarcity of observations for some of the typologies, TVP and Switching VARs were grouped into a single category. The remaining VAR model types are grouped into one category that includes VEC models.

Table 3: Types of VAR Methodology found in the Literature by Categories

▪ Time-varying Parameters (TVP) and Switching VAR	▪ Vector Auto-regressive (VAR)
- TVP-VAR	- VAR
- TVP-VAR with Stochastic Volatility	- Vector Error Corrected (VEC)
- TVP-FAVAR	- Recursive VAR
- MSVAR	- Recursive VAR with dummy
- MS-FAVAR	- Signed-restricted VAR
- Regime-switching SVAR	- Structural VAR
	- Non-linear VAR

[8] This group, named *Other Timeframes*, comprises the first two programs - First QE program and the CME, all three - First QE program, CME, and the qualitative and quantitative easing (QQE), solely the CME, when the period coincides with the comprehensive monetary easing program, solely the QQE - the same for the qualitative and quantitative easing program; and CME/QQE.

▪ Bayesian inference – Bayesian VAR

- *Variables that measure output and other variables used in the regressions* - The next set of data, presented in Table 4, includes all of the variables used in each model described in the sample from which the estimated impulse responses were produced.

Table 4: List of Variables that Compose the Models Reported in the Literature

<table>
<tr><th colspan="6">A) Categories of Variables and their Respective Sub-categories:</th></tr>
<tr><td colspan="2">Variable(s) that measure(s) the output: GDP; GDP growth; and Output gap;</td><td colspan="2">Monetary Variable: Monetary Base; Money Stock</td><td colspan="2">Secondary Monetary Variable: M2; M3</td></tr>
<tr><td colspan="2">Price level (or Proxy): CPI; Interest Rate; Core CPI Inflation Gap; GDP Deflator</td><td colspan="2">Interest Rate of Reference: Call Rate; 3-month interest rate; Repo Rate</td><td colspan="2">Exchange Rate: Nominal Yen/Dollar Spot Rate; Nominal Effective Exchange Rate; Real Effective Exchange Rate; Trade</td></tr>
<tr><td colspan="3">Spread: Difference between the 5-year JGB yield and the Call Rate; Difference between the 10-year JGB yield and the Call Rate</td><td colspan="3">Bond Yield: 10-year JGB yield or JGB yields</td></tr>
<tr><th colspan="6">B) Synthesized Variables (= 1 if used; = 0 if not)</th></tr>
<tr><td colspan="6">Stock Prices (or Stock Price Index); Bank of Japan Stock Purchases; Bank of Japan Bond</td></tr>
<tr><th colspan="6">C) Other Variables (= 1 if used; = 0 if not)</th></tr>
<tr><td colspan="3">Oil Inflation Rate; Bank of Japan ETFs Purchases; Bank of Japan J-REITs Purchases; Non-performing Loans in Japan; Japanese Exports; Government Expenditure; Commodity Price; Loans and Discounts in the Japanese Banking System; Bank Lending in Japan; Bank Share Prices; Condo Price Index; Average Lending Rate (on loans and discounts with maturity of less than one year at the time of origination)</td><td colspan="3">Value of Civil Engineering Projects (government expenditure); Interest Rate Factor (applicable to FAVAR models only); Price Level Factor (applicable to FAVAR models only); Yield Level Factor; Yield Slope Factor; Yield Curvature Factor; Gini Coefficient of Income Inequality; Dummy Variables; CPI inflation of Energy and Food (Exogenous Variable); Indirect Observance of Bank of Japan Monetary Policy</td></tr>
</table>

The first set (A) is composed of variables that vary greatly from study to study. To synthesize the collected information these were discriminated by category and we fit them into sub-categories[9], allowing a more parsimonious comparison. Additionally, similar variables were synthesized into broader concepts (subset B in Table 4), which accounts for synthetic variables built by researchers intending to measure the BoJ's policy stance over time.[10] The third subset of variables (C) comprises those that are not differentiated into subsets, since they do not belong to any specific category. In the first group, *monetary base* is a broad sub-category that includes not only the estimates that used the monetary base but also others that used one of its sub-components (BoJ's outstanding current account balance, CAB, or reserve balance or ratio). In the same way, the *money stock* also accounts for any of its sub-components, which in the present case appear in the form of Japanese government bonds (sub-component of the L category of the Japanese broad money stock concept). In addition, the *Secondary Monetary Variable* includes only the aggregates M2 and M3, which were also included in models that already had a first monetary variable. In opposition to the variables in *Monetary Variable*, these aggregates were not regarded as monetary policy tools.

On the features of the IRF and shocks, we retrieved the following:

- *IRF window (in months)* – we distinguish short-term (until 25 months), medium-term (between 26 and 48 months), and long-term (more than 48 months).
- *Type of shock (1)* – the initial goal was to characterize the disturbance in the monetary policy variable in three ways: which variable was hit by the disturbance, the technique used to produce the disturbance, and its magnitude. A considerable portion of the shocks is not for the usual 1% or one standard deviation (SD) increase in a given monetary policy (MP) variable. In addition, authors often test several different policy tools, e.g., call rate and/or a money stock proxy for a period under money supply targeting. Therefore, we first characterize the shocks and then aggregate those by the affected variable (see Table 5). Positive or percentage increase shocks are commonly applied to the call rate, unless the time of the analysis comprehended ZIRP periods, for which some authors opted for other short-term interest rates such as the 3-month rate or the repo rate. More related to the quantitative easing itself are the shocks reported to a *money stock* or *money supply targeting*, or to *financial operations engaged by the BoJ*. Shocks to BoJ's current account balance or average outstanding account balance (AOAB) are a straightforward reflection of QE operations that affect the banking system reserves and then the money stock, before hitting output, and shocks to bank reserves (or reserve rates) are an indirect consequence of QE

[9] The full list of variables registered from the literature selection and subsequent designated category or sub-category (when applicable) is available upon request.

[10] To illustrate, the synthesized variable - *Stock prices (or Stock price index)*, is a tag for variables that we do not see the need to differentiate. For this particular case, they are stock prices: Tokyo stock price index, NIKKEI stock prices, and NIKKEI average stock price index.

operations that will eventually affect money in circulation. In its turn, when authors apply a shock to a money stock they are assessing the effect in the last stage and how it will affect output. On the other hand, authors also try to relate the impact of financial operations directly related with the large scale asset purchase program by assessing the effect of government bonds purchases in the output. Shocks that could not be categorized with the above-mentioned elements, and that do not fit in any of the latter types of shock, were registered as “Other Types of Shock”. The criterion used to group these types of shocks follows the one in the *Monetary Variable* category. The shock to a short-term interest rate of reference (SSTIRR) includes the shock to the call rate and its proxies. The shock to the money stock (SMS) includes the shocks to the JGBs. The shock to the monetary base (SMB) includes the shocks to CABs, AOABs, reserves, and reserves rate. Table 5 summarizes these types of shocks and the groups to which they belong.

Table 5: Type of Shocks found in the Literature

Interest-rate: Shock to the Call Rate – SSTIRR; Shock to the Short-term Interest Rate – SSTIRR
Money stock or money supply targeting: Shock to the Money Stock – SMS; Shock to the Current Account Balance – SMB; Shock to the Average Outstanding Account Balance (AOAB) – SMB; Shock to the Reserves – SMB; Shock to the Reserves rate – SMB
Financial Operations engaged by the BoJ: Shock to (Japanese) Government Bonds – SMS
Inflation: Shock to the Core CPI Inflation – OTS
Other Types of Shock – OTS

1) Broader categories: SSTIRR – Shock to Short-term Interest Rate of Reference; SMS – Shock to the Money Stock (or sub-component); SMB – Shock to the Monetary Base (or sub-component); OTS – Other Type of Shock.

- *Confidence intervals* (CIs) – the impulse response functions found in the selected literature are often accompanied by their corresponding confidence intervals. There are three approaches used to define these intervals. Many studies define the intervals in terms of percentage/percentiles – 95%, 90%, and 68% – whereas other studies define them in terms of standard deviations. The most common – ±one- and ±two-standard deviation confidence intervals – are roughly equivalent to 68% and 95% confidence intervals, respectively. There are also studies that use percentiles, which have been registered in two sets – 10th and 90th percentile confidence intervals (PCI) and the 16th and 84th. Primiceri (2005) is often cited as mentioning that under the assumption of normality, the 16th and 84th confidence intervals correspond to a ±one-standard deviation confidence interval. Our database registers the 68%, 90%, and 95% intervals and estimates with 10th and 90th PCI were coupled with the 90% CI in a single group. In opposition to Rusnak *et al.* (2013), estimates computed without confidence intervals were not excluded from the meta-analysis, thus allowing for comparisons between a broader and a narrower sample.

- *Are the output and monetary policy variables in levels or in first differences?* – This is in order to explain the occurrence of explosive behaviours in impulse responses.
- *The date of the beginning of the shock (if applicable)* – this permits us to identify the studies in which there are impulse response functions that are set to affect a concrete period in time and in which the disturbance coincides with at least one of the QE programs, e.g., a shock set to affect the data starting at 2002:Q1. This variable is also useful to frame TVP-VAR impulse responses. Because we are only interested in the impulse responses that were computed during QE periods, we considered, whenever available, only the periods of 2002-06 (first QE program[11]), 2010-11 (CME), and 2014-(…) (QQE). For example, Kimura and Nakajima (2016) analyse the period 1981:Q2 - 2012:Q3, for which they report all impulse responses twelve months after the initial shock. In this case, we registered for 2001:Q4, the magnitude of the shock in the 4th quarter. This procedure was replicated for selected QE time units available in this study's timeframe, one year apart from each other: 2002:Q4, 2003:Q4, 2004:Q4, 2005:Q4, 2009:Q4, and 20010:Q4.

3.2.4 Estimates

The following set of information is composed of elements taken from the observation of the impulse response functions. These elements seek to summarize if, and how, monetary policy shocks have been affecting output. These elements were registered only for impulse response functions that were previously acknowledged and registered as statistically significant.

- *Signal of the shock's impact on the output variable* – The intention here was to capture the overall effect of the monetary policy shock on output. It was registered if the effect in the output variable is mainly positive or negative, given that the impulse response is statistically significant at some period of its length. We registered as "1", the significantly positive QE shocks; significantly positive accommodative non-QE shocks; and the negative contractionary non-QE shocks. Thus, "1" stands for the estimates' sign that supports the notion that the monetary policy tool increased output (in absolute terms). "-1" was used to register the opposite results and also whenever the registered value was null. "0" was used to mark all the non-statistically significant estimates.
- *Persistence of the Shock's Impact on Output* – In VAR-type models when the shock has a statistically significant impact on output it lasts for a certain period until it fades and disappears. We measure this period in months.
- *Magnitude: value of the shock's impact on the output variable (in percentage intervals)* – the variation of output has been registered for publication bias screening assessment purposes

[11] Because the first QE starts only in March 2001 and we are registering the shocks at the beginning of the year, we do not consider the year 2001 and start at 2002.

(Section 5) at 3rd, 12th, 24th, 36th, and 48th month after the shock's hit[12]; and has been registered at its maximum value when the impulse function has a statistically significant period[13]. Because the naked-eye observation of the figures provided in the studies is simply an imprecise technique to extract rigorous information, the values are displayed in intervals of magnitude, in order to reduce that level of imprecision. Nevertheless, it is most prudent to interpret this variable as an approximation indicator due to the impossibility of extracting the concrete values. The intervals are disposed as a cumulative sequence of 0.05%. In this way we hope that we are still able to provide a certain degree of detail among the collected estimates, given that the scales used in the literature to frame the dimension of the shock vary greatly, as much as from 0.01% to 1%.

3.3 Descriptive Statistics

In this section we analyse our database in detail. Table A.1 in the online Appendix A contains information about some of the variables analysed below, namely those being used from this point forward.

Regarding the authorial aspects of the sample, and in particular the year of publication, it is noticeable that the mean is approximately 2014, close to the upper bound of the observed range 2006 - 2020. Considering the full sample of 146 observations, an important portion (49.32%) of the research debate occurs outside the publishing sphere, composed of working papers and mimeos. The portion published in general journals is 43.15% and in the field monetary journals is 7.53%. It is worth mentioning that 14.38% of the observations are studies from which at least one author is (or was) directly associated with the Bank of Japan.

The data used in these studies do share some similarities. More specifically, 63.02% of the cases comprehend the years of the first QE program (2001-2006). This is understandable having in mind that at least 50% of the studies were published before 2014, but it also stresses the fact that less attention has been given to understanding the individual impact of the later QE programs. The most frequent periodicity is monthly data – 60.95% – followed by quarterly data – 37.67%. Around 47% of the studies use samples with over 100 observations, 37% have between 50 and 100 observations, and the remaining 16%, fewer than 50 observations.

Regarding the VAR approach used in each study, we mentioned above that there was a loss of accuracy by shorting down the list of the many VAR model variations found in the literature, in order that this information can become sufficiently parsimonious to be modelled within the meta-analysis context. Notwithstanding, we may add that within the TVPVAR-Switching group, the TVP approach (considering all its variations) is greater (32.20%) than the Switching VAR group

[12] These values were registered regardless of statistical validity. The moment zero has been registered separately as well, when available (contemporaneous shock).

[13] Marked as "NS" – Non-significant, if there are no statistically significant periods.

(8.89%). Similarly, entries in the database marked as the basic VAR account for 33.33% of the wider VAR-VEC group, which accounts for 57.53% of the whole sample.

Looking at the variables contained within those models, the ones most used besides output are in descending order of importance: monetary variable (92.47%), price level (89.05%), and interest rate of reference (47.95%); followed by the exchange rate (37.68%), stock prices (33.33%), and bond yield (31.11%). No surprises arise from these results, even though 33 different (categories of) variables were identified throughout the literature selection (besides the variable that measures output). The unmentioned 27 categories of variables are not of standard use, thus possessing little importance in the whole sample. Moreover, the number of variables in a model may vary between the minimum of 5 up to a maximum of 8. It may also be important to note that the type of shock that has been registered the most – shock to the monetary base (60.27%) – actually owes its weight to the shocks to the outstanding current account balance. The shock to the money stock excluding the shocks to the JGBs held by the BoJ account for 19.86 %; and expectably, the shock to the short-term interest rate of reference (10.95%) is mainly a reflection of the employment of shocks to the call rate.

At last we make a prior analysis of the information gathered for the Estimates section of the database, which correspond to the characterization of the monetary policy shocks to the (Japanese) output in terms of signal (overall effect) and magnitude. In terms of the overall effect of the shock, it is noticeable that a great portion of the estimates – 47.25% – suggest an overall significant positive effect of the BoJ's capability to promote the increase of output (to some undesignated extent); whereas an almost similar portion of estimates – 44.53% – did not find statistical proof to support that result. Furthermore, only a small portion of the estimates – 8.22% – indicate that the BoJ policies had a negative overall impact on output. in order not to detract from what has been reported in the literature selection, the (intervals of) magnitude of the behaviour of output were registered, and are shown separately, for categories that we previously established that differ according to the reasoning of inference and the type of monetary policy used – QE/Others and non-QE. Concerning solely the density of magnitude of QE/Other shocks to the output, what stands out is the large portion of studies that report a maximum statistically significant positive value of no more than 0.05%; all other intervals of maximum magnitude are relatively inexpressive if compared with this one. The expression and density of intervals is also of little importance, advocating that according to the theory, a shock to a QE monetary policy tool (the "Other Shocks" category has a small weight) is expected to affect the output positively, but in this specific case, with little or no relevance. Conversely, the same scenario may be traced for the non-QE estimates, despite an inexpressive landscape provided by a small number of observations.

4 Publication Bias Screening

In this section we assess the publication bias. This form of bias takes place whenever the results reported in the literature show evidence of patterns that are expected to occur in published studies. According to Stanley (2005, 2008), the quintessential forms of publication bias that may be found within a pool of collected estimates are of type I – the tendency for studies to report inflated results and/or the tendency for results to fall heavily to a single side of a central value – and of type II – the tendency for reported results to be statistically significant. These types of publication biases arise from decisions that researchers take at several stages in their work and are often difficult to distinguish from one another. Nevertheless, it is fairly accepted that authors are encouraged by peers and publishers to report strong and definitive evidence of whatever the subject is, in order to see their research published. In our own case, this would mean that if publication bias is detected, we should expect that there would be a bias to report significant effects of quantitative easing on output. This poses the selection bias or "file-drawer" problem as well, in which studies that mainly report statistically non-significant results and/or report "odd" results that do not comply with the established theoretical or empirical history are less likely to be published.

One way to analytically screen for publication bias within a pool of estimates, hereinafter also designated as effect sizes[14], is to relate the value of each estimate with a value that measures its estimation precision. In the absence of any systematic distortion, the expected relationship is: the greater the precision, the less variation around a "true effect". By "true effect" we mean an identifiable central value from which effect-sizes (of quantitative easing) may vary regardless of the level of variation (see, Begg and Berlin, 1988, and Stanley, 2001). This notion of true effect is important as a way to discern if a central value may be perceived as a proof that the relationship between quantitative easing policies and real output variables actually exists, assuming that the ultimate criterion applied is that there must be consensus among the literature. In that case, evidence of publication bias may manifest itself if the loss in precision (higher standard errors) is tied to the effect size value, because authors may report intentionally higher values to compensate for less precision. The analysis that follows is embedded in this latter idea and comprises two types of tests: the funnel asymmetry test (FAT) and the precision effect test (PET) (Stanley 2005, 2008; Doucouliagos and Stanley, 2009).

4.1 Funnel Asymmetry Test

The first part of the analysis is based on the scatter plots that put effect sizes against a measure of their statistical precision. Among the elements that characterize the impulse response functions,

[14] See, e.g., Kelley and Preacher (2012) for a comprehensive analysis on the concept of effect size.

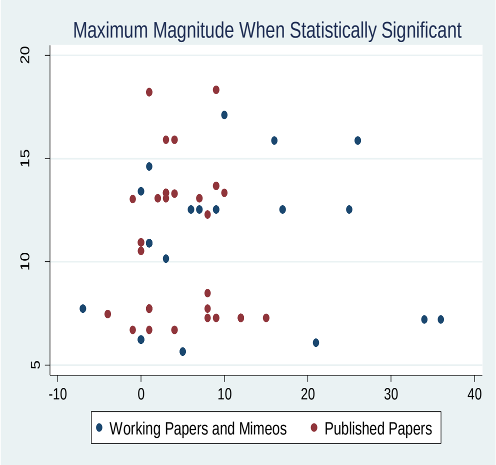
the magnitude of the response to the quantitative easing policy is the most suitable for measuring the effect size. As a measure of statistical precision, the literature on meta-analysis postulates the use of the inverse of the standard error (SE) of the estimated effect size. For this study the square root of the number of observations is used as an approximation of the standard error (Stanley, 2005), whilst being aware of some possible caveats. First, there is the assumption that the sample size is somewhat correlated with the measure of precision, which in this case is acceptable since the number of observations is in the formula of the standard error. Another required assumption is that the measure of precision is dependent on the sample size, but the inverse is not true, that is, the sample size is not fixed *a priori* to produce a certain level of variation around the estimate. This assumption holds in this study because there is no evidence that the studies chosen had pre-determined sample sizes, something unusual in this type of literature but rather common in experimental frameworks. We exclude the entries within the TVP-Switching category since in these cases the reported timeframe of analysis may not match the sample size used to produce the model's parameters estimates. Rusnák *et al.* (2013) uses a distinct standard error source by taking the magnitude length that goes from the value of the effect size to the corresponding confidence band that is closer to the horizontal axis. Instead, the data and methods we use are more heterogeneous and it is harder to obtain the band values. For example, some confidence bands that were provided are not immediately convertible into a single measure, such as the 90 PCBs and the 10th/90th percentile CBs.

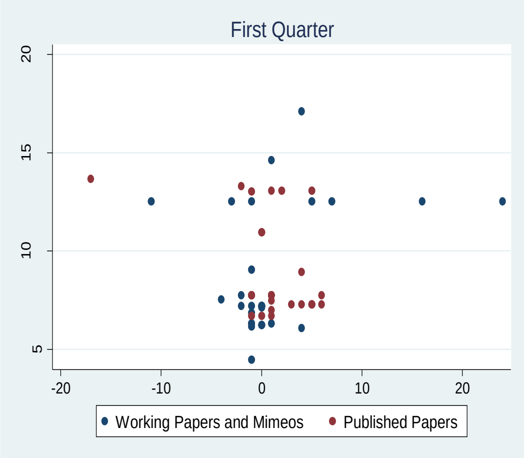
When analysing this type of scatter plot, the absence of systematic bias should give us a very clear perception of a relationship between the effect size and its measure of precision, in the sense that as the latter increases, the variation around a "true effect" size should diminish evenly. This relationship should appear, by visual inspection, as a pyramidal shape of the scatter plot or, as it was coined in the literature, an inverted funnel shape. The detection of type I publication bias is based on a plot feature that, besides having a funnel shape, is not symmetrical, i.e., the effect sizes tend to vary with greater incidence to the right or left of the central value. One further problem here is that it would not be prudent to analyse effect sizes that differ in terms of the monetary policy tool used. It is a legitimate concern to expect that non-QE tools, such as the call rate, may affect output in a different order of magnitude when compared with QE tools, e.g., increase of the current account balances. In consonance with this idea, Stanley (2005) observes that a wide number of different methodological and data specifications across studies may result in different true effects and may force the plot distribution to be skewed to one side without the presence of publication bias. All things considered, our analysis includes subsamples taken from the original 146 observations and the funnel plots displayed concern only QE (and Other Shocks) sizes: the maximum magnitude, when statistically significant, and the magnitude values at the 1st quarter, and 12th and 24th months horizons (see Figure 1 for each of the four cases). We also distinguish between the effects presented in published papers and those of working papers and mimeos.

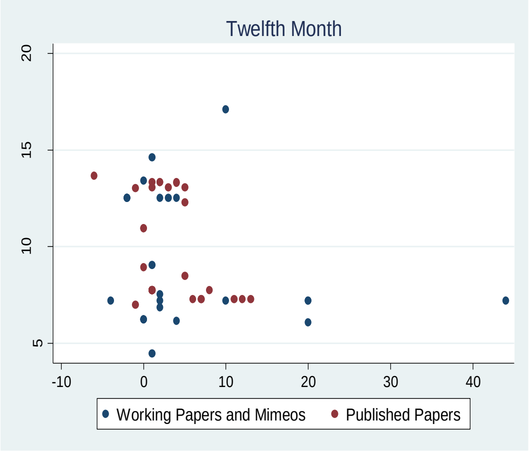
When observing the scatter plots, we notice a high concentration around a small interval of magnitude values, with the exception of the maximum magnitude when statistically significant effect sizes[15], where this is less clear. This indicates that there might be a consensus in the literature around what can be designated as the true effect of QE in output variables. In the maximum magnitude plot we detect a funnel shape whether we are considering published and non-published papers together or separately. In addition, we can notice that non-published effect sizes are more disperse than published ones, with special emphasis to the right side of the plot. Regarding first quarter effect sizes, the funnel shape looks less prominent than the former plot, although the concentration of values within an interval of magnitude is more visible (between -5 and 5 approximately). It is also noticeable in this plot that non-published effect sizes are more concentrated in negative territory (right next to zero), whereas published papers appear more concentrated close to zero but positive. Twelfth month and 24th month plots show similarities in the sense that both funnels' shapes look more evident compared to the former plots. Regardless of being published or unpublished, effect sizes are more disperse to the right of the main concentration of values. Also in both cases, with slightly more emphasis for the 24th month plot, there is more dispersion of non-published effect sizes on the right side when compared to published ones, i.e. non-published studies reported higher magnitude values. As we have seen earlier, the observation of a funnel shape with a thicker or skewed side indicates the presence of publication bias, although the plots here presented do not constitute definitive evidence of such. Unfortunately, the lack of observations for non-QE effect sizes leaves us almost nothing to work with. Since observing just these specific plots might be insufficient for detecting patterns that lead us to infer possible publication bias, we ought to conduct a second form of analysis that may help to shed some light on this issue.

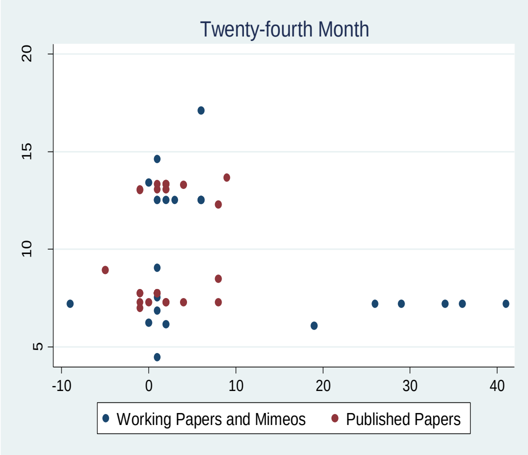
Figure 1: Effect-sizes Collected from the Literature Selection

[15] Hereinafter abbreviated as maximum magnitude effect sizes.









4.2 Precision Effect Test

A usual complementary approach to the funnel analysis is to fit a linear regression in order to assess the possible statistical relationship between the effect sizes and their precision. The regression is given by:

$$\widehat{\gamma_\iota} = \beta_0 + \beta_1 Se_i + \varepsilon_i, \qquad i = 1, \dots, n, \qquad \varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2), \tag{1}$$

where the i^{th} estimated size effect (of QE in real output variables), $\widehat{\gamma_\iota}$, depends on its standard error, Se_i. Here, a true fixed value effect is given by β_0 and publication bias is absent whenever β_1 is statistically non-significant. Otherwise there could be some publication bias. The error term ε_i, is normally distributed and independent across i. Consider the alternative model in which Se_i is set equal to the inverse of$\sqrt{n_i}$:

$$t_i \equiv \frac{\widehat{\gamma_\iota}}{Se_\iota} = \beta_1 + \frac{\beta_0}{Se_i} + \varphi_i \Leftrightarrow \widehat{\gamma_\iota}\left(\sqrt{n_i}\right) = \beta_1 + \beta_0\left(\sqrt{n_i}\right) + \varphi_i, \tag{2}$$

$$\varphi_i \backslash \sqrt{n_i} \sim \mathrm{N}(0, \sigma_\xi^2).$$

To account for heteroscedasticity, (2) becomes the "weighted version" of the first specification in which the dependent variable, $\widehat{\gamma_\iota}\left(\sqrt{n_i}\right)$, is now a proxy of the t-statistic and the measure of precision is no longer inverted (Stanley, 2005). Now, consider the following model:

$$t_{ij} = \beta_1 + \beta_0\left(\sqrt{n_{ij}}\right) + \epsilon_{ij} + \alpha_j, \quad \alpha_j\backslash\sqrt{n_{ij}} \sim \mathrm{N}(0, \delta_\alpha^2), \qquad (3)$$

$$\epsilon_{ij}\backslash\sqrt{n_{ij}} \sim \mathrm{N}(0, \delta_\epsilon^2), \quad i = 1, \ldots, m_i. \;\; j = 1, \ldots, m_j.$$

The specification (3) is a generalization of the preceding version that accounts for the possibility that the effect sizes are correlated to some extent within the same study[16], due to methodological and/or data similarities. This type of model[17] is designated as multi-level mixed-effect (MLME), which can be seen as an extension of the simple linear regression models (OLS). The observations are now discriminated as the ι_{th} result inside the j_{th} study. The model is called mixed because it includes a "fixed" part as in (2) and a "random" part given by $\epsilon_{ij} + \alpha_j$, in which we assume that α_j embodies a measure of within-cluster correlation. More so, this is said to be random because it represents the variance at our specific cluster level and does not depend on the "fixed part" of the model. Also, $Var\left(\varphi_{ij}\right) = \delta_\alpha^2 + \delta_\epsilon^2$ and given that $Cov\left(\varphi_{ij}, \varphi_{gj}\right) = \delta_\alpha^2, i \neq g$, the intra-class correlation (ICC) between individual and cluster level is $\rho_\varphi = Cor\left(\varphi_{ij}, \varphi_{gj}\right) = \delta_\alpha^2/(\delta_\alpha^2 + \delta_\epsilon^2)$ and assumes that the within-cluster error is equally correlated (Cameron and Miller, 2015). In our study we compare the MLME model's results with the analogous versions of OLS with cluster-robust standard errors[18], which may be seen as an alternative approach to cluster modelling. In addition, whenever there is no clear evidence that the MLME approach provides a better fit than the simple OLS, we study this last specification.

4.3. Estimation Results

Table 6 presents model (3) estimations. The exercise is performed for all studies and for published studies only. First, we look at the results gathered from the effect sizes registered at its maximum value during a statistically significant period. The LR test results show a similar scenario both when all studies are considered and just the published ones. In both cases the MLME models are preferred over the simple pooled OLS models and the ICC values are above 0.5, reinforcing the idea that the cluster specification of the models is relevant. Moreover, the results reveal that for both scenarios the constant parameter is not statistically significant, which suggests that there is no publication or systemic bias. The intercept parameter β_1 is statistically significant at a 10% level only when only published studies are considered and at the maximum and the 1st quarter

[16] An alternative specification could be to cluster by author instead of by study.

[17] Also designated as hierarchical model since it belongs to the type of model that allows accounting for the correlation between groups of observations, in which clusters composed of the initial set of observations may be nested at a second (smaller) level of clusters.

[18] Hereinafter designated as cluster OLS. This model specification is intended to prevent over-rejection of the parameter's statistical significance (Cameron and Miller, 2015).

effects. This suggests that for short horizons the published studies induce the existence of a possibly small publication bias. Regarding true effects, mixed results are obtained: we obtained significant true effects for the 1st quarter and maximum magnitude (one negative and significant when considering published papers) but we also obtained almost non-significant true effects when considering longer time spans (12 and 24 months). Regarding the 1st quarter effect sizes, when all studies are considered we notice that the cluster specification does not seem suitable (ICC is zero), i.e., there is no within-study correlation among observations. The simple pooled OLS model confirms all the previous conclusions: no true effect and no bias.

As a matter of fact, these results may suggest that QE policies in Japan may have no effect on average in real output variables, in both the very short (1st quarter) and more medium runs (24th months), i.e., Japan may not rely on those policies to relaunch the economy when in crisis. Moreover, this may help to explain why, despite the long and persistent QE approach of the BoJ, the Japanese economy has achieved only sluggish economic growth rates in the past few decades (0.2% yearly between 2000 and 2019).

Table 6: Main PET Results

	Maximum Magnitude When Significant		1st Quarter		12th Month		24th Month	
	MLME – All studies	MLME – Published Studies	OLS – All Studies (1)	MLME – Published Studies	MLME – All studies	MLME – Published Studies	MLME – All studies	MLME – Published Studies
β_1 (Intercept)	-13.282 (42.760)	-20.768 (34.866)	-12.267 (25.090)	121.549* (64.973)	27.830 (33.012)	25.425 (37.694)	15.199 (43.639)	-57.287 (40.085)
$\beta_0(\sqrt{n_{ij}})$	6.801* (3.511)	5.675** (2.834)	2.567 (2.755)	-13.34** (6.428)	0.217 (3.093)	-0.680 (3.417)	2.129 (3.961)	7.645* (3.524)
LR test (p-value)	0.000***	0.000***	-	0.000***	0.001***	0.000***	0.000***	0.000***
ICC ratio	0.737	0.776	-	0.968	0.344	0.867	0.474	0.895
Obs..	66	40	66	33	60	35	55	32
N. Clusters	26	16	-	12	19	12	17	10

Notes: *, **, *** if statistically significant at 10%, 5%, and 1% confidence level, respectively. Parameter's standard errors in parentheses. The null hypothesis of the LR test states that the MLME model is equivalent to a simple pooled linear regression model. ICC ratio ranges from [0;1], the correlation within clusters. OLS instead of the MLME version, since in this latter version the ICC is 0.

With the exception of published studies evaluating the maximum and the 1st quarter effects, taking the wider view regarding the last analysis, the first conclusion is that there is a sense that the literature does not provide consensus on what might be a true effect, since heterogeneity of methodologies and frameworks across studies may give rise to several true effects. Furthermore, when comparing the results from both FAT and PET, it becomes clear in all cases that the concentration of effect sizes around a smaller interval of magnitude is insufficient to assume the existence of a true effect, and what seemed a possible bias from the funnel plots was not corroborated by the PET results.

To further screen for possible publication bias, we extended the analysis adding a dummy variable to the models that comprehend all the effect sizes (published and unpublished). The dummy takes the value of 1 if the effect size is reported in a study whose author is associated with the Bank of Japan. According to the methodology applied by Doucouliagos and Stanley (2009), the variables added to model (3), intended to screen for publication bias, are now jointly interpreted with the intercept, and the overall bias is the net value of their sum. From all the results obtained[19], there is no evidence that studies produced by authors related to the Bank of Japan report biased effect sizes. This evidence applies for all the sets of effect sizes with the exception of maximum magnitude, which is more ambiguous. An investigation related with this topic is Fabo *et al.* (2020), using 54 studies for the UK, the US, and the Eurozone. The authors analyse the effect of QE on output and inflation in those three economies, while also taking into account other characteristics of the studies, including biographical information of the authors of the studies (i.e., if each study had at least one author from a central bank). With this last information the authors were assessing if there were any differences between results obtained by academics *versus* results written by researchers working at a central bank. Results of their study show that research made by central banks' staff report larger (and significant) effects of QE on output and inflation than research made by academics. These results seem to be driven by career concerns and are more pronounced amongst senior researchers. The share of studies that included at least one author working at a central bank is around 60%. In our case the share of studies that were made by (at least) an author(s) working at a central bank is 14.4%, a much smaller share than the 60% in Fabo *et al.* (2020).

[19] The models with dummy include all the sets of effect sizes – maximum magnitude, 1st quarter, 12th month, and 24th month. The results are available upon request.

5. Meta-Probit Estimation

In a meta-analysis it is usual to extend the PET model by including as explanatory variables several elements that characterize and differentiate the studies from which the effect sizes are taken, seeking thereby to understand whether there is a significant relationship between the magnitude of the average effect size and each of those elements. Thus, in this Section we search for possible additional determinants of the effect size of quantitative easing policies on output in Japan observed in primary studies. Because we found difficulties in measuring the precise value of the collected IRF's magnitudes, we chose to perform a type of meta-analysis based on a probit regression. That is, we are able to use a larger dataset in our meta-analysis by reading from the studies' plots where the IRF's lie (positive or negative responses) and from their conclusions in terms of the Granger causality tests. This type of regression uses the same set of explanatory variables included in the extended PET model to answer a different kind of question: *does a given study characteristic, e.g., the type of output variable used in the study's framework, affect the probability of it reporting an (overall) sign and significance effect size, based on a monetary policy shock?* As we can see, this question no longer uses a real-valued dependent variable, but a categorical one, which includes the information that we gathered and termed as overall response sign of the real output and the statistical significance of explanatory variables, namely those linked with the unconventional monetary policy. An important detail can be reported back to the descriptive statistics: because there are so few overall IRF negative effects registered (8.22% of the total), we perform not an ordered probit, which would segment the dependent variable into positive-significant effect, insignificant effect, and negative-significant effect, but a standard probit model in which insignificant and negative-significant fall into the same category. The positive effects are 47.25% and the non-significant 44.53%, in the 146 observations. Thus, the dependent variable can be read as a positive and significant effect if "1" ("success") or "0" otherwise (as in Kluve, 2016).

5.1 Further Data Transformation for the Meta-Probit Analysis

An important concern when constructing categorical variables is to avoid using (sub-)categories that are not representative in the sample. Related to that, some of these control variables present too many categories that would translate into many dummy variables and reduce the number of degrees of freedom in the model. To avoid these problems, we proceeded to another round of variable's re-categorization. There is no clear-cut justification for when to consider a category's representation small, so variables with a sub-category with 10% or fewer observations were subject to re-categorization[20].

The first variable is the Journal Impact, which ranks the effect sizes into three categories – A, B, and C – resorting to the SCIMAGO econometric journal impact ranking, based on the idea presented in Koetse *et al.* (2009). If an effect size is taken from a published study classified in the SCIMAGO ranking within the first two quartiles – Q1 or Q2 – it receives a classification "A". It is ranked "B" if the study has one of the SCIMAGO bottom classifications – Q3 or Q4 – or if it has not been considered in SCIMAGO. Finally, the effect size ranks "C" if it was taken from non-published studies. With this ranking variable we seek to assess whether journals with greater impact tend, or not, to report positive significant estimates (a form of publication bias screening). Our database comprises 26.7% of studies published in A journals, 32.3% of studies published in B journals, and 41.1% of studies are non-published studies (the C category). The *Number of Observations of the Analysis' Timeframe* is used as a quantitative variable instead of having two dummies and a reference group. 46.6% of the studies use more than 100 observations, while 37% use between 50 and 100 observations, and the remaining 16.4% use fewer than 50 observations. The number of observations *per se* may not be suitable for a posterior marginal analysis of the probit model. Alternatively, we create a variable in which the number of observations falls under ordered intervals of 20 observations. Each of these variables is used one at a time in the probit model. The *Variable that Measures the Output* also has too many sub-categories, which were reduced in two different ways, to be used alternately: the first specification distinguishes variables that are in levels from those that are in differences (reference group); in which 75.2% are in levels and 24.8% are in differences. The second specification distinguishes industrial output variables (59.9% of the database) from all other forms (40.1%). The *Monetary Variable* is used in the model as originally intended, distinguishing money base (87%) from money supply variables (5.5%) and NA cases (7.5%), but will also alternate in the probit model with the specification that distinguishes solely the money base against "others" (13% - money supply or NA). In the *Types of Shock* variable, the smallest categories – Other Type of Shock and Shock to a Short-term Interest Rate – are now Other Types of Shock (19.9%), with the drawback of removing interest rates shocks from the inferential conclusions. *Confidence Intervals* categories that do not correspond to 1-SD width were aggregated into one category (45.2%). Furthermore, the Interest Rate of Reference (48% of studies use this variable) and Price Level (89% use this one) are included in the model in simply a binary fashion, "present or not present", instead of discriminating every sub-category with dummies. Finally, effect sizes based on daily observations are joined with monthly-based effect sizes (62.3% of the database), and in the Empirical Methods,

[20] The exceptions are the SCIMAGO based ranking variable (sub-category rank A with 10.27%) and the monetary base variable that includes the money supply category (5.47%), both used in the first stage of the probit analysis.

Bayesian VAR effect sizes are added to the larger TVP-VAR and Switching VAR group (42.5% of the studies). A summary of the variables that need further treatment is in Table A.2 in the online Appendix A.

5.2 The Probit Model and Preliminary Procedures

The specification of the probit model is the following:

$$Pr(y_i = 1 | x_i) = \Phi(x_i b) \tag{4}$$

$i = 0, 1, 2, \ldots,$ n.

$y_i = 1 \equiv$ statistically significant positive effect size

where Φ is the cumulative distribution function (CDF) of a standard normally distributed random variable and $x_i b$ is a linear combination of the explanatory variables – the index function. The explanatory variables, x, are a set of studies' characteristics; the binary outcome y is "1" if the effect size is statistically significant and positive, "0" otherwise (negative or not significant); b gives the model's coefficients.

Because we gathered a great number of possible explanatory variables, we decided for a modelling strategy based on the fundamentals proposed by Hosmer and Lemeshow (2000, p. 116), in order to convey a structural sense to the analysis. The use of this modelling strategy is intended not only to narrow down the number of variables included in the model, but also to build the model in a stepwise basis that may allow for more statistically robust model estimations. We start with a preliminary analysis of the variables to be included in the model. One variable at a time or a categorical group as a whole (Hosmer and Lemeshow, 2000, p. 38), we include it in the probit model and examine its significance by resorting to the p-values of the coefficient's Z-statistic (which in this case is also the model's Wald statistic). We refer to this process as Univariate analysis (UNIVARA). This approach was performed twice for each variable: a standard probit model and a version of it with robust standard errors, in which the data were clustered by the paper of origin. Because the coefficients' values can change when interacting with other variables, we exclude variables at this stage only when both standard and cluster estimations report p-values over 0.1. Table A.3 in the online Appendix A reports the p-values of the coefficient's Z-statistic of the univariate models.

There is a group of variables that was excluded ("Ex") from the next stage of the analysis. The variables excluded were: the studies' impact ranking; whether an author is associated with the Bank of Japan; whether a VAR/VEC type of model was used instead of a TVP-VAR; Bayesian VAR or Switching VAR; whether the output variable is measured in levels or in first differences; and the inclusion of an interest rate variable, an exchange rate variable (37.7% of the studies in the database), or a bond variable in the model. As a result, these features are not relevant to the sign and significance of the effect size of QE on output for the Japanese economy.

In an alternative to the original two-dummy scheme, the monetary variable (MV1) was tested in a single dummy form – "1" if money supply, "0" otherwise – but the result turned out to be non-significant. Therefore, we considered the two-variable scheme in the next stage. The category that discriminates the type of shock applied to the output variable is statistically significant in both standard and clustered forms when the reference group is the monetary base or other variables rather than the monetary base or the money supply. The results from the individual tests did not perform as well: the dummy variable that discriminates shocks applied to the monetary base (Shock a) violates the 0.1 threshold criterion and the dummy variable that discriminates shocks other than the monetary base or money supply (Shock c) has similar p-values. In addition, the individual tests related to the use of different types of confidence intervals were found to be highly non-significant when discriminating 1 SD confidence intervals (CI1), which may suggest that the probability of success does not depend on this standard band width and when discriminating other SD and confidence bands (CI2).

5.3 Stepwise Variable Selection and the Estimated Probit Model

The next stage of the analysis uses a stepwise forward variable selection with backwards elimination as part of the model's modelling strategy. Its purpose is related with the fact that after the UNIVARA stage there are still many variables to add into the probit model and there is no theory that supports a hierarchy of relevance regarding what variables should be chosen over others. The list of variables is the one previously cleared at the UNIVARA stage. For the Stata algorithm, the significance level used for the coefficient's Z-statistic was set at 0.08, a value that allows the inclusion of borderline variables that exceed the conventional 0.05 level. After exhausting the tests of inclusion, the algorithm starts a search backwards for the removal of the least significant coefficient until all variables are significant. The initial specification starts with the addition of the variable that performs best given the 0.08 criterion. As in the UNIVARA analysis, the stepwise estimation process was performed using standard errors and cluster robust standard errors.

The final estimations results are shown in Table 7. The selected results take into consideration the performance of the Pearson's Chi-squared goodness-of-fit (GoF) test, which infers the ability to fit the probit model well to the data. But because the stepwise algorithm does not take into account how good a model is in terms of its GoF, we performed a subsequent round of estimations that consisted in removing or replacing one particular variable at a time so that there is an improvement of the GoF of a model and/or reaching alternative model specifications. Finally, because "Year" is a trend-type variable, we also considered different specifications by including or not this variable in the models. The models we considered are: the benchmark models 1-a (with Year) and 1-b (without Year); models 2-a and 2-b, which follow from 1-a and 1-b, respectively, after not rejecting the equality of the coefficients associated with MV1 and MV2[21]; and models 3 and 4, in which the variables Periodicity, Year, CI1, CI2, Shock_a, Obs (and Int. Obs) were removed, substantially increasing the GoF tests.

For a matter of robustness check and to somehow summarize the estimation results obtained from the six abovementioned models, we also considered model averaging smoothing schemes for maximum likelihood estimators (last column of Table 7). See the monograph by Claeskens and Hjort (2008). For each coefficient, the first value averages all six models, the second one those that include Year (1-a and 2-a), and the third entry averages the models that do not include Year (1-b, 2-b, 3, and 4). The model's weights follow from the smoothed-BIC (SBIC) formula (the SAIC are very similar). When all models are considered, it is 1-a (0.43) and 2-a (0.56) that obtain almost full weight. When we average models 1-b, 2-b, 3, and 4, the weight rank goes as: 2-b (0.44), 1-b (0.38), 4 (0.18), and 3 (0.00). Therefore, the top two model averaging estimates are very similar but different from the third averaging model selection.

From a general perspective, the estimation results presented in Table 7 report standard, clustered, and cluster bootstrap[22] SEs that are not overwhelmingly high. According to Hosmer and Lemeshow (2000, pp. 135-141), suspiciously high coefficient values and/or standard errors may be caused by the numerical problems earlier enunciated: quasi- or complete separation and collinearity between variables. In terms of the selected variables, adding Obs or Int. Obs in the models did not produce any notable estimations, and thus we can infer that more observations do not affect the probability of attaining a positive and statistically significant effect size. The same happens when the effect size is controlled for shock c, i.e., if the shock is applied to other variables besides monetary base and money supply.

On the other hand, in model 3 with standard SEs, we conclude that published papers have a higher probability of reporting a positive significant size effect. Published papers represent 50.7% of the database. There is evidence in models 2-b and 4 that studies focused on the first Japanese quantitative period (2001-2006) have a lower probability than other QE periods to find a positive and statistically significant output effect. The studies that focus on only the first QE period represent 63% of the total. Models 1 and 2 suggest that studies that use (daily or) monthly data over quarterly data have an increased probability of producing positive and significant effect

[21] From model 1-a we obtained the p-values of 0.0991 for testing MV1=MV2 and 0.0010 for CI1=CI2. From model 1-b we got 0.3289 for testing MV1=MV2 and 0.0084 for CI1=CI2.

[22] We used the boottest stata command with the option bootcluster(code) proposed by Roodman *et al.* (2019). We considered the score bootstrap (Kline and Santos, 2012) as an adaptation of the wild bootstrap because as they explain, "In estimators such as probit and logit, residuals are not well-defined, which prevents application of the wild bootstrap."

sizes, hence higher frequencies help in finding a significant positive effect. Moreover, we find that the most recent studies have a lower probability of finding a positive and significant size effect. Model 4 suggests that if a study uses industrial output as its explanatory variable (output proxy) the probability of estimating a positive and significant result is greater compared to the use of other output variables. In this same model (albeit a weak statistical significance) including the price level as an explanatory variable is more likely to produce a positive and significant effect size than not using it. Also, both *MV1* and *MV2* variables, or when merged into a single one, have coefficients with negative signs. This suggests that regardless of the type of monetary variable used – money supply (MV1) or money base (MV2) – including it in the model results in a lower probability of producing a significant and positive effect size than using other kinds of non-monetary variables. On the contrary, both CI1 (one SD) and CI2 (other than one SD) have point estimates with positive signs in which the first the larger, thus suggesting that having confidence intervals with one standard deviation implies a higher probability of producing a significant and positive effect size compared to models that use other types of confidence intervals. Finally, from the estimations, only the dummy variable that distinguishes shocks to the monetary base (shock a) from any other type of shock has been included in three of the reported models (1-a, 2-a, and 2-b). There is evidence that the use of a monetary base tool increases the probability of producing a positive and significant effect size.

The model averaging point estimates confirm all of these coefficients' signs. To wrap up, the variables that have a greater probability of producing a significant and positive effect size are: published papers, studying more than just the first Japanese QE period, using monthly (and daily) data, older studies, using industrial output as the output proxy, including the price level but not a monetary variable, one SD confidence intervals, and having a monetary base type of shock.

Table 7: Meta-Probit Estimation Results

	Model 1-a	Model 1-b	Model 2-a	Model 2-b	Model 3	Model 4	Model MA
Publication	-	-	-	-	0.528 (0.221)** (0.367) [0.169]	-	0.000 0.000 0.000
QE		-0.636 (0.280)** (0.399) [0.108]	-0.502 (0.292)* (0.360) [0.184]	-0.657 (0.284)** (0.363)* [0.091]*	-0.412 (0.227)* (0.332) [0.229]	-0.760 (0.270)*** (0.339)** [0.026]**	-0.287 -0.286 -0.529
Periodicity	- 0.767 (0.253)*** (0.306)** [0.029]**	-0.852 (0.251)*** (0.317)*** [0.021]**	-0.779 (0.252)*** (0.299)*** [0.023]**	-0.552 (0.271)** (0.336) [0.109]	-	-	-0.773 -0.774 -0.564
Year	-0.162 (0.045)*** (0.080)** [0.005]***	-	-0.107 (0.037)*** (0.045)** [0.037]**	-	-	-	-0.130 -0.131 0.000

Industrial Output (IO)	-	-	-	0.434 (0.261)* (0.347) [0.288]	0.412 (0.220)* (0.356) [0.245]	0.594 (0.230)*** (0.308)* [0.066]*	0.001 0.000 0.191
Price level	-	-	-	-	-	0.677 (0.389)* (0.513) [0.308]	0.000 0.000 0.000
MV1	-2.173 (0.611)*** (0.437)*** [0.013]**	-1.699 (0.550)*** (0.350)*** [0.018]**	-1.978 (0.582)*** (0.436)*** [0.014]**	-2.122 (0.565)*** (0.413)*** [0.002]***	-1.251 (0.464)*** (0.466)*** [0.004]***	-1.281 (0.467)*** (0.504)** [0.008]***	-2.060 -2.062 -1.573
MV2	- 1.667 (0.613)*** (0. 620)*** [0.071]*	-1.936 (0.558)*** (0.443)*** [0.007]***	-1.978 (0.582)*** (0.436)*** [0.014]**	-2.122 (0.565)*** (0.413)*** [0.002]***	-1.251 (0.464)*** (0.466)*** [0.004]***	-1.281 (0.467)*** (0.504)** [0.008]***	-1.843 -1.844 -1.663

CI1	1.895 (0.472)*** (0.555)*** [0.007]***	1.850 (0.449)*** (0.609)*** [0.007]***	1.845 (0.452)*** (0.596)*** [0.009]***	1.927 (0.453)*** (0.607)*** [0.003]***	-	0.913 (0.265)*** (0.331)*** [0.014]**	1.865 1.867 1.545
CI2	0.876 (0.470)* (0.565) [0.191]	1.097 (0.438)** (0.605)* [0.194]	0.818 (0.458)* (0.586) [0.310]	1.176 (0.443)*** (0.576)** [0.115]	-	-	0.843 0.843 0.931
Shock a	0.582 (0.271)** (0.320)* [0.083]*	-	0.534 (0.265)** (0.310)* [0.123]	0.442 (0.260)* (0.308) [0.204]	-	-	0.553 0.555 0.194
Constant	326.942 (90.678)*** (163.083)**	0.843 (0.532) (0.572)	216.357 (75.300)*** (91.285)**	0.454 (0.566) (0.576)	0.747 (0.462) (0.492)	0.075 (0.608) (0.782)	262.922 264.021 0.517
GoF’s *p*-value(1)	0.0043	0.0000	0.0000	0.0025	0.4426	0.1293	-

LR test Homo. vs Hetero.	0.0103	-	-	-	0.4085	0.4673	-

Notes (1) GoF’s p-value is taken from Pearson’s Chi-square; (2) Top parentheses are s.e. for the standard model; Bottom parentheses are cluster robust s.e. at the study level (45 clusters were used); Square brackets are the p-values from the cluster bootstrap t-tests; (3) ***,**,* are statistical significance at 1%, 5%, and 10%, respectively; (4) LR test Homoscedasticity versus Heteroscedasticity is the Stata’s test – “Likelihood-ratio Test of lnsigma^2 = 0” – which checks the overall significance of the regressors on the probit’s variance (more details below in the subsection “Heteroscedasticity Analysis and Interaction Effects”. For convergence, we included in the heteroscedasticity equation the variables QE, MV1, MV2, CI1, and CI2; (5) The description of the variables can be found in the online Appendix A and the “model MA” is a model average of the main 6 models, as explained in the main text.

5.3 Estimated Marginal Effects

For a better understanding of the above results, we obtain and report the marginal effects extracted from the estimated models in Table 7. This excludes model 1-a since our previous findings suggest that it may be heteroscedastic. The marginal effect is the change in the predicted probability of success associated with the change from "0" to "1" for a dummy and setting all other variables at a given value. In this case we use the actual observed values of the other variables and calculate averages, an approach designated as average marginal effects (AME). For example, suppose we want to know the change in the predicted probit probability of success when the price level is considered (used in the model that produces the output estimation) in opposition to not being used. First, calculate the predicted probability of the "first" effect size, that is, by setting the price level as not in the model, "0" and using the estimated probit and the actual observed values of the other variables in x_i with $i = 1$.. Repeat in the same manner for the other 145 observations. Then take the average. Second, proceed as before but now assuming that the price level is used, "1", in all 146 cases and then take the average of those predicted probabilities. Third, compute the overall average marginal effect of using a price level variable by subtracting the average predicted probability of not using the price level from the average predicted probability of using it.

That is, these two scenarios produce different predicted probabilities and the difference is our (averaged) marginal effect.

Table 8 presents the estimated AME. One way to analyse these effects is to compare their values within the same model and then across models. The AMEs with the greatest magnitude are those of MV1 and MV2. This suggests that whenever in a study either a monetary supply variable (MV1) or a money base variable (MV2) is chosen over other types of monetary variable, the probability that the study finds a positive and significant effect size decreases immensely, roughly between 43% and 66%. A noticeable discrepancy is given by the different signs in MV and Shock a. In the first case, the predicted probability of using a money supply variable decreases, but if the shock used to produce the output estimations is of a money base type (*Shock a = 1*), then the probability increases between 14% and 16% (models 2-b and 2-a, respectively). One can then refer to the overall average marginal effect of choosing money base over other types of monetary variable as the difference of those two opposing effects, which is roughly between 30% and 50%. Because the monetary variable is a crucial one to produce the output estimations, it is not unreasonable to state that the type of monetary variable used affects the behaviour of the impulse response functions when a monetary policy shock exists, thus affecting the output estimation in terms of its sign and significance. This has direct policy implications to the variable chosen as a monetary policy instrument. The variable CI1 (more than CI2) also has high AMEs that close to the MV ones. Regarding the other variables, the magnitude of the AMEs for including a price level variable is greater than 22.5%, and thus more influential than the choice of an industry index as an output proxy (between 14% and 20%). The estimated AME for the specification that studies the 1st QE period indicate that more QE programs (that may or may not include the 1st QE) have to be included in the analysis, to increase the probability of reporting statistically significant (positive) effects of QE on output growth. Similarly, studies that make use of daily or monthly data increase the probability of reporting statistically significant (positive) effects of QE on output growth. The exception to the very high AMEs found is the case of the integer variable Year. Its AME is of a 3.1% decrease in the probability of studies reporting a positive and statistically significant effect size, if a study is more recent.

Table 8: Average Marginal Effects

	Model 1-b	Model 2-a	Model 2-b	Model 3	Model 4
Publication	-	-	-	0.186 (0.073)**	-
QE	-0.205 (0.086)**	-0.150 (0.085)*	-0.204 (0.084)**	-0.145 (0.077)*	-0.253 (0.083)***
Periodicity	-0.275 (0.071)***	-0.232 (0.067)***	-0.171 (0.080)**	-	-
Year	-	-0.031 (0.010)***	-	-	-
Industrial Output (IO)	-	-	0.135 (0.079)*	0.145 (0.075)*	0.197 (0.071)***
Price level	-	-	-	-	0.225 (0.125)*
MV1	-0.548 (0.161)***	-0.591 (0.154)***	-0.660 (0.150)***	-0.442 (0.152)***	-0.426 (0.144)***
MV2	-0.625 (0.160)***	-0.591 (0.154)***	-0.660 (0.150)***	-0.442 (0.152)***	-
CI1	0.597 (0.122)**	0.551 (0.115)***	0.599 (0.118)***	-	0.303 (0.078)*
CI2	0.354 (0.134)**	0.244 (0.133)*	0.365 (0.130)**	-	-
Shock a	-	0.159 (0.075)**	0.137 (0.078)*	-	-

Notes: 1) Delta-method standard-errors in parentheses. *,**,*** are statistical significance at 10%, 5%, and 1% level, respectively; (2) The description of the variables can be found in the online Appendix A and each of the five models includes the variables for which we present a point estimate.

5.4 Heteroscedasticity Analysis and Interaction Effects

5.4.1 Heteroscedasticity

A problem in the probit model (4) is that inconsistency of the MLE arises if heteroscedasticity is present (Davidson and MacKinnon, 1984). Heteroscedasticity can be defined as a non-constant variance that depends on the variable(s) in the index function. In the former probit model (4), homoscedasticity is assumed so that the normal CDF has a constant variance of 1, which may not be the actual case in practice. In fact, one could ask whether the probit estimations reported earlier in Table 7 are biased due to heteroscedasticity, or not. Stata provides a command that allows fitting a maximum-likelihood heteroscedastic probit model. This model yields a probability function of success of the following type:

$$Pr(y_i = 1|x_i, z_i) = \Phi\left\{\frac{x_i b}{exp(z_i \gamma)}\right\}, \qquad (5)$$

where, in the words of Harvey (1976), z_i "is a m *x 1* vector of observations on a set of variables which are usually, though not necessarily, related to the regressors x_i" of the index function; and γ "is a m *x 1* vector of parameters". For this analysis, we fit heteroscedastic probit models with the same variables as in the previous models reported in Table 7[23]. The estimation of this type of model is known to have difficulties in convergence and indeed it succeeded only for models 1-a, 3, and 4. For the remaining models (1-b, 2-a, and 2-b) no convergence was achieved after 16,000 iterations.

Resorting again to Table 7, the LR test for homoscedasticity at a 5% level pointed out that model 1-a presents evidence of heteroscedasticity, thus suggesting that this model specification is not reliable and might contain numerical problems. On the contrary, models 3 and 4 present an insignificant LR test, which indicates that there is no reason to suspect heteroscedastic bias at the previous standard versions, so their results are reliable. To confirm this fact, in the estimated model (5), none of the auxiliary model's variance coefficients γ are statistically significant.

5.4.2 Interaction Effects

The search for interaction effects among variables has the purpose of assessing whether the dependent variable is related to how two (or more) of its explanatory variables interact with each other. According to Norton *et al.* (2004), the interaction effects in non-linear models with two

[23] Regular standard errors were specified. The use of cluster robust standard errors was not considered because the Stata base manual refers to this option as inefficient for this type of model.

variables are not derived from marginal effects, as in the linear case, but from the cross-partial derivative of the expected value of y, so that we have for any X:

$$E[y|x_1, x_2, X] = \Phi(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + X\beta) = \Phi(u), \tag{6}$$

where β_{12} is the parameter from which the interaction effects between the two variables, in this case x_1 and x_2, can be derived. The interaction for non-linear models is given by the following cross-partial derivative:

$$\frac{\partial^2 \Phi(u)}{\partial x_1 \partial x_2} = \beta_{12}\Phi'(u) + (\beta_1 + \beta_{12}x_2)(\beta_2 + \beta_{12}x_1)\Phi''(u). \tag{7}$$

As the analysis of the interaction effects is somewhat burdensome, when considering all of the cases of cross-partial derivatives for all of the covariate values, we studied only the heteroscedasticity bias free models 3 and 4. First, we searched for the interaction terms that, regardless of their significance, do not inflict statistical non-significance on the other parameters. This was done by fitting one interaction term at a time[24]. Next, we analysed the interaction effects in terms of their statistical significance and then, conditionally to the previous point, their sign (and sign shifts) and the calculation of the cross partial derivatives.

From the previously described procedures, we found that the inclusion of the interaction term did not interfere with the signs and the statistical significance of the other parameters and that these interaction terms were not statistically significant.

6 Conclusion

We study the impact of the quantitative easing programs set by the BoJ on the Japanese output and/or output growth, by means of a meta-analysis, using a total of 146 observations, collected from 45 studies. We focused on the monetary policy literature that makes use of the vector auto-regressive methodology (in its various forms) and resort to the published impulse response functions that describe the output reaction to a monetary policy shock. The database built for the meta-analysis is a panel, containing elements that describe the estimated effect sizes and several other elements, both methodological and data-related. The main goal of the meta-analysis is to screen for biased reported results in published studies and understand whether elements that characterize the literature addressed affect the probability of studies reporting an overall positive and statistically significant effect of quantitative easing on output and/or output growth, by means of probit model specifications. Compared to other studies our probit specification is an innovative approach to analysing the effects of different features that contribute to the identification of a positive and significant effect of quantitative easing on output. Adding to the probit-meta

[24] We used Stata's command Inteff, i.e., fitting a model with one interaction term at a time.

approach, we also implemented new developments in meta-analysis; namely, the cluster bootstrap t-tests and the model averaging smoothing estimation schemes.

In our sample 47% of the effect sizes reported an overall positive and significant effect on output of the monetary policy measures implemented during a quantitative easing period; 45% were non-significant, and a residual number of *ca.* 8% were negative and significant.

Funnel asymmetry and precision effect tests do not provide overwhelming evidence of publication bias. Additionally, we also do not find a consensus regarding the output and/or output growth effects during the quantitative easing years. Furthermore, there is evidence that certain elements found in the framework of the literature addressed affect what is reported in terms of sign and significance of the response-functions. The meta-probit analysis suggests that industrial production and the price level have a greater effect on the probability of reporting statistically significant (positive) effects of quantitative easing on output and/or output growth. Different specifications in the data used, such as being an older study, if the study includes several QE programs, if it is published in a journal, if it is using monthly or daily data, and if it is using one standard deviation in the confidence intervals, also affect more the probability of reporting statistically significant (positive) effects of quantitative easing on output growth.

Additionally, if in a study either a monetary supply variable or a money base variable is chosen over other types of monetary variables, the probability that a study finds a positive and significant effect size decreases immensely. However, if the shock used to produce the output estimations is of a money base type, then the probability of reporting statistically significant (positive) effects of quantitative easing on output growth increases. Because the monetary variable is a crucial one to produce the output estimations, it is not unreasonable to assert that the type of monetary variable used affects the behaviour of the impulse response functions when a monetary policy shock exists, thus affecting the output estimation in terms of its sign and significance. This has direct policy implications regarding the variable chosen as a monetary policy instrument.

When compared with previous review studies (e.g., Papadomou *et al.*, 2019; Papadamou *et al.*, 2020; Poter and Smets, 2019), our results seem to suggest a more sceptical view of the real effect of unconventional monetary policy, since the average effect is sometimes non-significant and different specifications lead to different quantitative true effects. One of the most important (policy) implications of our study is that until now the evidence for a significant real effect of unconventional monetary policy is weak, since concerning true effects, we obtain mixed results. Moreover, this may help to explain why, despite the long and persistent QE approach of the BoJ, the Japanese economy has showed only sluggish economic growth rates in recent decades (0.2% yearly between 2000 and 2019).

In future research it would be interesting to extend this analysis to other central banks that have adopted unconventional monetary policies, i.e., quantitative easing, to see if the results are perhaps country specific.

References

1. Bhattarai, S. and Neely, C. 2018. An Analysis of the Literature on International nconventional Monetary Policy, *Working Paper 2016-021C, Federal Reserve Bank of St. Louis* https://doi.org/10.20955/wp.2016.021
2. Begg, C. B. and Berlin, J. A. 1988. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society*, (Series A) 151: 419–445.
3. Buch, C., Bussiere, M., Goldberg, L., and Hills, R. 2019. The International Transmission of Monetary Policy. *Journal of International Money and Finance* 91, 29-48.
4. Cameron, A. C. and Miller, D. L. 2015. A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50(2): 317.
5. Claeskens, G. and Hjort, N. L. 2008. *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.
6. Davidson, R. and MacKinnon, J. G. 1984. Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics*, *25*(3): 241–262.
7. Dell'Ariccia, G, Rabanal, P., and Damiano, S. 2018. Unconventional Monetary Policies in the Euro Area, Japan, and the United Kingdom. *Journal of Economic Perspectives*, 32 (4): 147-72. DOI: 10.1257/jep.32.4.147
8. Dekle, R. and Hamada, K. 2015. Japanese monetary policy and international spillovers. *Journal of International Money and Finance*, 52: 175-199.
9. Doucouliagos, H. and Stanley, T. D. 2009. Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations*, 47(2): 406.
10. Fabo, B., Jancoková, M., Kempf, E., and Pastor, L. (2020). Fifty Shades of QE: Conflicts of Interest in Economic Research. *Working Paper No. 2020-18*, Becker Friedman Institute for Economics at University of Chicago.
11. Fatouh, M. (2016). Post Crisis Unconventional Monetary Policy in the UK, the US, and the EA, *Mimeo*, Department of Economics, University of Essex.
12. Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The Next Generation of the Penn World Table. *American Economic Review*, 105(10), 3150-3182, available for download at www.ggdc.net/pwt.
13. Gambacorta, L., Hofmann, B., and Peersman, G. (2014). The Effectiveness of Unconventional Monetary Policy at the Zero Lower Bound: A Cross-Country Analysis. *Journal of Money, Credit and Banking*, 46: 615-642. doi:10.1111/jmcb.12119
14. Ganelli, G. and Tawk, N. (2019). Spillovers from Japan's Unconventional Monetary Policy: A Global VAR approach. *Economic Modelling* 77, 147-163.
15. Gros, D., Alcidi, C., and De Groen, W. (2015). Lessons from Quantitative Easing: Much ado about so little? *CEPS Policy Brief nº 330.*

16. Grauwe, P. D. and Storti, C. C. (2004). The effects of monetary policy: A meta-analysis. CESifo Working Paper No. 1224.
17. Harvey, A. C. (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, 44(3): 461-465.
18. Havranek, T. and Rusnak, M. (2013). Transmission lags of monetary policy: A meta-analysis. *International Journal of Central Banking*, 9(4): 39-75.
19. Havránek, T., Stanley, T. D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., and Rost, K. (2020). Reporting Guidelines for Meta-Analysis in Economics. *Journal of Economic Surveys*, 34 (3): 469-475.
20. Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression. 2nd ed.* 392 p. Wiley Series in Probability and Statistics – Applied Probability and Statistics Section. John Wiley and Sons, Inc., Hoboken, NJ, USA.
21. Kelley, K. and Preacher, J. K. (2012). On Effect Size. *Psychological Methods*, 17(2): 137–152.
22. Kimura, T. and Nakajima, J. (2016). Identifying conventional and unconventional monetary policy shocks: A latent threshold approach. *B. E. Journal of Macroeconomics*, 16(1): 277-300.
23. Kluve, J. (2016). A Review of the Effectiveness of Active Labour Market Programmes with a Focus on Latin America and the Caribbean. International Labour Office. Research department working paper no. 9.
24. Kline, P. and Santos, A. (2012). A score based approach to wild bootstrap inference. *Journal of Econometric Methods* 1(1): 23-41.
25. Koetse, M. J., Groot, H. L. F. De, and Florax, R. J. G. M. (2009). A Meta-Analysis of the Investment-Uncertainty Relationship, *Southern Economic Journal*, 76(1): 283–306.
26. Lombardi, D., Siklos, P., and St. Amand, S. (2018). A survey of the international evidence and lessons learned about unconventional monetary policies: a ‘new normal’ in our future? *Journal of Economic Surveys*, 32: 1229-1256. doi:10.1111/joes.12293
27. Montgomery, H., and Volz, U. (2019). The effects of Unconventional Monetary Policy in Japan. *Journal of Economic Issues*. 55: 411-416.
28. Norton, E. C., Wang, H., and Ai, C. (2004). Computing Interaction Effects and Standard Errors in Logit and Probit Models. *The Stata Journal*, *4*(2): 154–167.
29. Papadamou, S., Kyriazis, A., N., and Tzeremes, G., P. (2019). Unconventional monetary policy effects on output and inflation: A meta-analysis, *International Review of Financial Analysis*, 61, 295-305.
30. Papadamou, S., Siriopoulos, C., and Kyriazis, N.A. (202). A survey of empirical findings on unconventional central bank policies, *Journal of Economic Studies*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JES-04-2019-0186

31. Poter, S. and Smets, F. (2019). Unconventional monetary policy tools: a cross-country analysis, *Bank of International Settlements, CGFS Papers No 63.*
32. Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies,* 72(3): 821-852.
33. Ridhwan, M. M., de Groot, H. L. F., and Nijkamp, P. (2010). The impact of monetary policy on economic activity – evidence from a meta-analysis. Tinbergen Institute, Tinbergen Institute Discussion Papers: 10-043/3.
34. Roodman, D., MacKinnon,J., Nielsen, M., and Webb, M. (2019). Fast and wild: bootstrap inference in Stata using boottest. *Stata Journal* 19(1): 4-60.
35. Rusnak, M., Havranek, T., and Horvath, R. (2013). How to solve the price puzzle? A meta-analysis. *Journal of Money, Credit, and Banking*, 45(1): 37-70.
36. Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *The Journal of Economic Perspectives*, 15(3): 131-150.
37. Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics*, 70(1): 103.
38. Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys*, 19(3): 309-345.
39. Turner, P., Rabucci, A., Zhou, J., and Monnin P. (2019). The Risks and Side Effects of UMP: An Assessment of IMF Views and Analysis, IEO Background paper, May 14, 2019.
40. Ugai, H. (2015). The transmission channels and welfare implications of unconventional monetary easing policy in Japan. TCER Working Paper E-102. Hitotsubashi University. Kunitachi.
41. Uusküla, L. and Pitzel, D. (2007). The Effect of Financial Depth on Monetary Transmission. *The IUP Journal of Monetary Economics*. 5(2): 63-73.