ISCTE Business School
Instituto Universitário de Lisboa

A META-ANALYSIS ON THE BANK OF JAPAN QUANTITATIVE EASING POLICY: THE BANK OF JAPAN'S EFFECTIVENESS TO PROMOTE ECONOMIC GROWTH

Pedro André Simões Roque Linhares

Dissertation submitted as partial requirement for the conferral of

Master in Economics

Supervisor:

Prof. Alexandra Ferreira-Lopes, Assistant Professor, ISCTE Business School, Economics Department

Co-supervisor:

Prof. Tiago Miguel Guterres Neves Sequeira, Associate Professor, Universidade da Beira Interior, Management and Economics Department

September 2018

– Spine –

A META-ANALYSIS ON THE BANK OF JAPAN QUANTITATIVE EASING POLICY: THE BANK OF JAPAN'S EFFECTIVENESS TO PROMOTE ECONOMIC GROWTH

Pedro André Simões Roque Linhares

ISCTE Business School
Instituto Universitário de Lisboa

Abstract

As the Qualitative and Quantitative Easing programmes are still in place there have been signs that the Japanese economy will maintain a path of moderate economic growth, still, without glancing the desired 2% inflation. The question over whether and how the Bank of Japan successive quantitative easing programmes, which were based on an unprecedented increase on the central bank's asset sheet, have been successful in promoting a steady growth of the Japanese Economy, has been debated by the literature that is focused on the transmission channels of monetary policy. We present a comprehensive meta-analysis that focus on the literature that have been studying the effectiveness of the Bank of Japan's policies during the 2001 to 2016 period, that resorts to the Vector Auto-regressive methodology to analyze, through impulse response functions, how monetary policy shocks impact output. An analysis based on funnel plots – Funnel Asymmetry Test – and linear regressions – Precision Effect Test – does not provide evidence of publication bias, neither the consensus over the output growth during the quantitative years. A meta-probit analysis suggests that a study with the characteristics mentioned above, which uses certain variables to build the model – industrial output, price level, bond yield and either the money base or the money supply – as well as different specifications in the data used – increasing the number of observations used or choosing quarterly data – will affect the probability of reporting statistically significant output growth; notwithstanding, the evidence found in this last analysis varies in terms of statistical robustness.

Keywords: Quantitative easing, Bank of Japan, Effects of monetary policy on economic growth

JEL Codes: E52, E58

Resumo

Ainda com o programa de *Qualitative* e *Quantitative Easing* em vigor, têm existido sinais de que a economia japonesa manterá um caminho de moderada recuperação económica; não obstante, sem se vislumbrar o desejado crescimento da inflação a 2%. A questão em torno de se, e como, os sucessivos programas de *quantitative easing* baseados num crescimento sem precedentes dos ativos do Banco Central do Japão, têm tido sucesso em promover o crescimento estável da economia japonesa, tem sido discutida na literatura que se foca nos mecanismos de transmissão da política monetária. Neste estudo, apresentamos uma meta-análise que se foca na literatura que estuda a eficácia das políticas do Banco do Japão durante o período de 2001 a 2016. Literatura essa que recorre a metodologia baseada em modelos *Vector Auto-regressive*, para analisar através de funções de resposta a impulso, como é que os choques causados por ferramentas de política monetária afetam a produção da economia japonesa. Com base numa análise em gráficos de dispersão em funil – *Funnel Asymmetry Test* – e em regressões lineares – *Precision Effect Test* – não obtivemos provas que sugerissem *publication bias* – enviesamento dos resultados publicados em revistas – nem provas que sugerissem um consenso entre a literatura visada, relativamente ao valor do crescimento da atividade económica no Japão durante os períodos de *quantitative easing*. Uma análise baseada em modelos *meta-probit*, sugere que a inclusão, em estudos com a estrutura atrás mencionada, de certas variáveis no modelo a estimar (relativas à economia Japonesa) – o *output* industrial, o nível dos preços, as taxas de retorno de títulos da dívida japonesa, ou tanto a base monetária como a oferta de moeda nacional – tal como outras especificações relativas ao tipo de dados utilizados – o incremento do número de observações ou a utilização de dados trimestrais – podem afetar a probabilidade das estimações virem a reportar um crescimento positivo e estatisticamente significativo na atividade económica. Os resultados encontrados nesta última análise variam em termos de robustez estatística.

Keywords: *Quantitative easing*, Banco Central do Japão, Efeitos da política monetária no crescimento económico

JEL Codes: E52, E58

Acknowledgements

A particular acknowledgement of my gratitude to Professor Alexandra Ferreira Lopes and Professor Luís Martins for the time and the patience.

Table of Contents

List of Acronyms

3MIR – 3-month interest rate

AME – Average Marginal Effects

AOAB – Average Outstanding Account Balance

APP – Asset Purchase Program

BBS – Bank Balance Sheets

BMA – Bayesian Model Averaging

BoJ – Bank of Japan

BOJPGB – Bank of Japan Purchase of Government Bonds

BOJSP – Bank of Japan Stock Purchases

BVAR – Bayesian Vector Auto-regressive

CAB – Current Account Balance

CCIG – Core CPI inflation Gap

CCPI – Core Consumer Price Index

CDF – Cumulative Distribution Function

CEC – Credit Easing Channel

CGPI – Corporate Good Price Index

CI – Confidence Intervals

CME – Comprehensive Monetary Easing

CPI – Consumer Price Index

CR – Call Rate

DEIT – Direct Effect of Inflation Targeting

ETF – Exchange Trade Fund

FAT – Funnel Asymmetry Test

FAVAR – Factor Augmented Vector Auto-regressive

FG – Forward Guidance

FIRSTDIF – First Differences

GDP – Gross Domestic Product

GDPD – GDP deflator

GJ – General Journal

HP filter – Hodrick–Prescott filter

ICC – Intra-class Correlation

IER – Increase in the Excess Reserves

Int. Obs – Intervals of Observations

IO – Industrial Output

IOBOJMP – Indirect Observance of Bank of Japan Monetary Policy

IR – Inflation Rate

IRF – Impulse Response Function

IRR – Interest Rate of Reference

IT – Information Technologies

JGB – Japanese Government Bond

J-REITS – Japanese Real Estate Investment Trust

LR – Likelihood Ratio

LTM – Latent Threshold Model

LVS – Levels

MB – Monetary Base

MJ – Monetary Journal

MLE – Maximum Likelihood Estimation

MLME – Multilevel Mixed Effect

MP – Monetary Policy

MPP – Monetary Policy Proxy

MS – Money Supply

MS-FAVAR– Markov Switching Factor Augmented Vector Auto-regressive

MSt – Money Stock

MSVAR – Markov Switching Vector Auto-regressive

MV – Money Variable

NA – Non-Available

NEER – Nominal Effective Exchange Rate

NS – Non-Significant

NYDSR – Nominal Yen/Dollar Spot Rate

Obs – Observations

OG – Output Gap

OLS – Ordinary Least Squares

OTF – Other Timeframes

OTS – Other Type of Shocks

PCI – Percentile Confidence Intervals

PET – Precision Effect Test

PL – Price Level

PP – Published Paper

PRE – Portfolio Re-balancing Effect

QE – Quantitative Easing

QQE – Qualitative Quantitative Easing

REER – Real Effective Exchange Rate

RIRC – Real Interest Rate Channel

RR – Repo Rate

S10YCR – Difference between the 10-year JGB yield and the Call Rate

S5YCR – Difference between the 5-year JGB yield and the Call Rate

SD – Standard Deviation

SE – Standard Error

SMB – Shock to the Monetary Base

SMS – Shock to the Money Stock

SP – Stock Prices or Stock Price Index

SPC – Stock Price Channel

SSTIRR – Shock to the Short-term Interest Rate of Reference

SVAR – Structural Vector Auto-regressive

TBBR – Total Bank Balance Reserves

TCU – Transmission Channel Undefined

TQE – Tobin's Q Effect

TVP-VAR – Time-Varying Parameter Vector Auto-regressive

TWREFER – Trade Weighted Real Effective Foreign Exchange Rate

UNIVARA – Univariate Analysis

UR – Unemployment Rate

VAR – Vector Auto-regressive

VEC – Vector Error Correction

WE – Wealth Effect

WP – Working Paper

ZIRP – Zero Interest Rate Policy

List of Figures

List of Tables

1. Introduction

The Bank of Japan has been employing measures of unconventional monetary policy for over 15 years now. A substantial empirical bulk of work has appeared since then on the effects of the measures of unconventional monetary policy on several fundamental variables of the Japanese economy. These measures were set in place, specially, to stimulate the anemic Japanese economy. However, the Japanese case is still perceived by the majority of the audience – policy makers, researchers, investors, or 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 in the opposite direction. In the following study it is analyzed the unconventional monetary policy of the Bank of Japan (BoJ) and its impact on the economic growth of the Japanese economy. Specifically, relying upon meta-analysis to withdraw valid and statistically relevant conclusions based on a selection of empirical literature that assesses and measures the impact of the behavior of the Bank of Japan in the Japanese economic activity, with particular emphasis on GDP and/or its growth.

This study has the following structure: Section 2 establishes the context of the study in regards to the Japanese case and introduces previous cases of meta-analysis literature, which focused on monetary policy transmission. Section 3 gives an account on how data has been collected from the literature on Japanese Monetary Policy, and how has been treated and organized. Section 4 presents descriptive statistics based on the dataset built and described in Section 3, giving an account on some key elements regarding the conclusions and estimations reported in that very same literature. Section 5 makes use of the same dataset to conduct a type of analysis based on funnel plots and linear regressions, to screen for biased results in the published literature here addressed. Section 6 presents a series of probit estimations that try to unveil 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.

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

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

The last 15 years of the Japanese economy can be described as a row of successive attempts to recover from the generally designated *Japan's lost decade*, the 1990's. This decade is considered a significant turning point for the Japanese economy, characterized by a long-lasting recession, which ended, in the beginning of the 2000's, in a combination of negative output gap (as well as sluggish economic growth) and moderate deflation. The prospects for the Japanese economy were many times clouded by the repercussions of internal financial crisis that stroke the economy over time: in the end of the 80's with the burst of the real estate bubble which contaminated all the financial system; the IT bubble that went bust in 2000; and more recently, the international financial crisis of 2008.

The efforts to revert the scenario persisted throughout the recent economic history of Japan, considered 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 revert the scenario of the 90's, the BoJ engaged in what was known to date to be an unconventional type of policy framework, substituting the main policy tools and the policy targets. The period that fall under this unconventional monetary policy approach, was designated as Quantitative Easing (QE). Since 2001 there have been three programmes that fall under the category of QE, being the last of the programmes implemented (and still in place), designated as the Qualitative and Quantitative Easing (QQE) programme[1]. There are several features in these programmes that may be pointed out as basic elements that form their identity. The first feature is the fact that this policy framework is called *a programme*: a conditional set of policy measures, for which pre-established rules determine their continuity or cessation. Adding to this point is the fact that a so called *programme* implies a sense of closure and goal achievement. This is an attempt by the BoJ to let know private agents of a more active intervention in the economic scenario; in opposition to a later accommodative stance during the 1990's. Finally, another feature that may be pointed out is the transition for more active

[1] For a small chronology of events see Annex I, Table 9-1.

operating mechanisms of monetary policy (MP), in opposition to the traditional policy tool, the overnight call rate[2]. Regarding the overnight call rate, the BoJ maintained what is called the Zero Interest Rate Policy (ZIRP) during the QE years[3], until it announced in late January 2016, for the first time, that the level of the call rate would be set at 0.1% below zero[4].

Using the definition of Ugai (2015), a QE programme can be seen as the work of two mechanisms that operate through both sides of a central bank's balance sheet (in this case the BoJ's). 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 is the liability side of the same balance sheet that is being used.

5.2 Monetary Policy Transmission under Meta-Analysis

In what regards the existing literature that could serve as a methodological object of comparison, one could only find a small number of studies that employ meta-analysis focusing on monetary policy transmission; and more specifically, that simultaneously distinguishes between types of VAR methodologies employed, and made use of an effect size based on impulse response functions[5] (IRFs). Notwithstanding, Table 2-1 resumes some of the existing literature addressed to other countries. For instance, Grauwe and Storti (2004) used a meta-regression to infer on 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. The same meta-analysis points out a large variation in the results reported in the literature, concerning the estimations for output; stating as well that part of that variation could be explained by whether the

[2] The call rate is the designated reference interest rate of the BoJ; an overnight interest rate that the BoJ uses in interbank operations.
[3] The ZIRP – when the overnight call rate was set between 1 and 0% – coincided with the QE frameworks during the periods of 2001 to 2006 and 2010 to 2016.
[4] Such novelty in the BoJ's monetary policy framework is not contemplated in any of the studies selected.
[5] The cited studies also used meta-analysis in order to unveil cross-country heterogeneity of results, which methodologically speaking, leads them to include in their regressions variables that distinguish estimates per country. These variables are based on economic features such has openness to foreign-trade or proxies for financial development. Regardless, our main purpose with this short review was only to stress some of the findings that relate the heterogeneity of the reported results with differences in approaches of the methodology.

authors used VAR or SVAR techniques. In a similar fashion, Ridhwan *et al.* (2010) conducted a meta-analysis, which reported an accentuated variance of estimated output effects taken from the literature; both regarding the speed and magnitude of transmission.

Pitzel and Uusküla (2007) found some evidence, for countries within the EU-15, that higher financial depth is positively correlated with a stronger transmission of monetary shocks. These authors took three exterior variables that measure financial depth for each country, and assessed their correlation 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, which aimed to capture the reasons behind the price puzzle patterns found across the literature. The findings in this study suggest that, often, patterns observed in the empirical estimates are not consonant with what the theory postulates; the authors also suggest that more observations exert a positive effect on the long-term estimates of the price level after a shock in the interest rate of reference (monetary policy tool), i.e. the price puzzle does not fade with time; and that the reported estimations do vary depending on the VAR specification and output proxy used.

Instead of employing a regression in their research, Havranek and Rusnak (2013) opted for a Bayesian model averaging (BMA) method, in order to withdraw 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 literature results. Moreover, these authors also found that data and methodology factors play a role in explaining the variation of results within the literature. Studies that use monthly data and report strictly decreasing impulse responses are prone to make 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.

Table 2-1: Effect Sizes used by Meta-analysis Studies on Monetary Policy Transmission Mechanisms and the Countries or Regions that were considered in the respective Literature Selection

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, US.
Pitzel and Uusküla (2007)	Maximum level attained by the monetary policy shock in output and price level.	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 level.	USA, *Eurozone* and European Union (*Non-Eurozone*).
Rusnak *et al.* (2013)	1% increase in the interest rate and in the price level caught at the 3rd, 6th 12th, 18th, and 36th month, plus 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, US.
Havranek and Rusnak	1% increase of the interest rate on price level caught at	(Same as Rusnak *et al.*

(2013)	the minimum level (after reaching its maximum) for *humped-shape responses*, and also at the last period available for strictly decreasing responses.	2013)

3. Data and Methodology

A two-stage process was applied to create a pool of studies, which begun with the search for all the studies available in Google Scholar, RePEC, B-on, and Scopus, possessing the following features:

- Attempted to respond to the question whether the Japanese Monetary Policy was effective in promoting economic growth during the QE periods, even if this was not the main question.
- Made use of Vector Auto-regressive (VAR) methodology (or related methodologies).
- And the methodological framework supported their statistical results with the use of impulse response functions (IRFs).

By imposing these features to every study selected, we account for a certain degree of homogeneity within that pool, thus creating the necessary basis for comparability between studies. As an additional criterion, regarding the impulse responses functions, these must account for a shock caused by a monetary policy tool that conveys its impact onto an output proxy. The graphical representations of these impulse responses are the source from which was possible to extract several important features about the literature on the given subject, e.g.: are the monetary policy shocks affecting positively or negatively the Japanese economy?, what is the magnitude and duration of such impacts?, can the authors identify transmission channels through which those impacts are conveyed?, what can one say about the statistical robustness of these IRFs?

Having these features in mind, several combinations of the following set of words were typed in the 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 and much more successful moment of the search process was *snowballing* from the results found in the first moment[6]. The search was conducted between March and May 2016.

The data was collected from 30 studies[7], registering a total amount of 104 impulse response estimations, covering a publishing period from 2006 to 2016, being the most recent publications considered preferably, since the length of the QE programmes covered were bigger. One criterion settled was that in order to be eligible, a study should include in its period of analysis roughly one year of quantitative easing. A direct consequence of this was to exclude studies or estimates based in periods of analysis before 2002. Aside from this, it was further decided that there would not be a prior exclusion of studies based on their publication characteristics such as their *status* – the prior expectation that for whatever reason, e.g., the reputation of the author(s) or the journal in which the study is published, might be a source of prior discrimination in the meta-analysis, by distributing more weight to the reported estimations perceived to be more trustable/reliable – or impact – the attempt to quantify that status, e.g., a study's number of citations within a given period. Despite of not having considered initially these two elements, a treatment of this nature will be given and described in Section 6.

3.1 Methodology - Construction of the Database

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. With this database, we intended to register all occurrences of this type in the literature selection here presented. Because, often, the information reported in impulse response functions is not quantified and summarized in a systematic manner, one had to withdraw it from the graphical

[6] *Snowballing* is to continuously look for the citations found in studies that are, or may be important, and to go look into the cited papers to search for what studies they have cited. Reverse *snowballing* was also performed for every study added to the selection – instead of looking for what a study cites, one searched in *Google Scholar* and the other platforms for what studies have cited a given study.

[7] The following studies, even though excluded from the literature selection due to no compliance with the selection criteria – do not present impulse responses with intervals of confidence –, are relevant for the discussion within the literature regarding the QE efficacy in the Japanese output: Kamada and Sugo (2006); Kimura *et al.* (2002); Nakajima *et al.* (2010); and Nakajima (2011a).

representations available. For such process, a miter was recurrently used for support, albeit there are numerous factors that pose hindrances to its good practice[8]:

- Often, authors are slightly loose on their judgment, when confirming the statistical significance of their estimates in borderline scenarios, based on the intervals of confidence.
- Graphical representations are often poor in quality; something as having a thick grid or more detailed axes' scales could help to visualize results in a more precise manner; nevertheless, seldom, was this observed as current practice.
- Was also uncommon to find in the literature, systematic information on summary statistics such as maximum and minimum values; the number of statistical significant periods, etc.; thus, relegating to the reader some portion of interpretation.

We have excluded the impulse responses that are presented in eventual robustness sections of the literature. These latter exercises tend to support or validate the researcher's main conclusions and to include them could create a bias on the results found on the meta-regression, once the extension of these robustness checks vary from study to study. Estimates were also excluded when the researcher(s) presented results but disregarded them in the first place, has being irrelevant and/or justifying their computation just to support a preliminary premise. Estimates produced via data simulation, Panel VAR or estimates reported in 3D representations were also excluded[9]. Moreover, the withdrawn estimates per study were not restricted to a fixed number, to prevent further selection bias.

3.2 Description of the Database

The content withdrawn from the study selection is here systematized in a database that serves the production of summary statistics and later on, the meta-regression. There are four broad groups of information – *Authorial Information*, *Data*, *Methodological Specifications* and *Estimates* (see Table 3-1). The construction of the database evolved,

[8] Rusnak *et al.* (2013) went a step further in good practicing by contacting the authors when in doubt about the graphical representation of the IRFs.
[9] The only Panel VAR study that we came across, Gambacorta *et al.* (2012), was not eligible according to our criteria. On the other hand, impulse responses depicted in 3D graphics were just too inappropriate to accurately collect the estimations.

first, the collection of rawer and more detailed data which in a second step was aggregated into broader categories of information. This was required in order to permit a certain level of statistical consistency (preventing the excessive reducing of the degrees of freedom) during the estimation of the meta-regression.

Table 3-1: Resume of the Information Collected to Build the Database

Authorial Information	Data
- Authors - Year of Publication - Type of Publication - Are the Author(s) associated with the Bank of Japan?	- QE Programs comprehended in the Analysis' Timeframe - Periodicity of the Time Series - Number of Observations of the Analysis' Timeframe - Midpoint of the Study's Timeframe
Methodological specifications	**Estimates**
- Empirical Method - Variable(s) that measure Output - Other Variables used in the Regression - IRF Window in Months - Type of Shock (1) - Confidence Intervals - Are the Output and Monetary Policy Variable in Levels or in First Differences? - The Date of the beginning of the Shock (if applicable)	- Statistical Validity of the Impulse Response based on the Granger Causality. - Signal of the Shock's Impact in the Output Variable - Accumulated Effect of the Shock's Impact on the Output Variable - Persistence of the Shock's Impact in the Output Variable - Magnitude: Value of the Shock's Impact in the Output Variable - Transmission Channels Thought to Affect Output

3.2.1 Authorial Information

Authors – the information regarding each impulse response function is identified by its study of origin.

Year of Publication – we registered the most recent publication date of each study, known at the time of the search period.

Type of Publication – the collected studies were either: published articles, working papers, or mimeos. For further investigation on publication bias, this selection of studies has been differentiated in two ways: the first distinguishes the published papers from working papers (including mimeos here); the second, published papers are distinguished between general and those specialized in monetary themes.

Are the Author(s) associated with the Bank of Japan? – being the Bank of Japan the central bank that officially dictates the monetary policy, by distinguishing the studies which are under the support of this institution, we may proceed with another publication bias screening: comparing the results found on the literature between the group of studies which are and are not associated with this policy maker. In this regard, this variable presents itself as a simple “yes/no” dichotomy.

3.2.2 Data

QE Programmes comprehended in the Analysis' Timeframe – because the underlying subject of analysis is the effect of the three known QE programmes on output, the comparisons between entries must account for the fact that different timeframes are used for several reasons; these may depend on the data of publication, restriction to the availability of data, or the desire of the researcher to study a period that comprises specific events, e.g. choosing a time frame that may comprise only one or more periods under different quantitative easing programmes. To alleviate the problem lifted by the existence of many timeframes we chose to group them in the following way: one group is composed by the studies that analyze a timeframe that only comprehends the first QE programme; the other group is composed by studies that do not analyze exclusively the first QE programme, or that analyze the other programmes[10]. The justification to aggregate the information in such way came from the need to make our data parsimonious and suitable to econometric modeling, and also in this specific case, to account for the fact that the attention given to each of the mentioned timeframes is highly uneven. Expectably, due to a greater time distancing, a large portion of the studies analyze the first QE programme alone, whereas, the other timeframes were much less used.

Periodicity of the Time Series – this variable will permit to assess a basic hint on the preference of the researchers regarding the periodicity of the data of choice. The periodicities registered were daily, monthly, and quarterly.

Number of Observations of the Analysis' Timeframe – there is an obvious correlation between the periodicity of the timeframe and the length of the timeframe, i.e. quarterly data may provide shorter time series compared with monthly data, and subsequently shorter time series than daily data. Moreover, the exact number of observations from which the impulse responses are estimated is sometimes omitted by the authors, which sometimes provide only an approximate number, or only the date at the beginning and at the end, from which the time series length is extracted. Based on these constraints, this variable is solely an approximation of the number of observations used for each entry (estimated by the date limits provided in each study)**.** Furthermore, the information has been labeled in the following way: lower than 50 obs.; between 50 and 100 obs.; and higher than 100 obs.

Midpoint of the Study's Timeframe – albeit not used in the estimations, it has been registered for sake completeness in Table 9-2, Annex I.

3.2.3 Methodological Specifications

[10] This last group, named *Other Timeframes*, comprises all the studies that include in their timeframe of analysis the first two programmes – First QE programme and the CME; all three – First QE programme, CME and QQE; solely the CME – when the period analyzed coincides with the Comprehensive Monetary Easing programme; solely the QQE – the same for the Qualitative and Quantitative Easing programme; and CME/QQE – when data's timeframe comprises these two programmes.

Empirical Method – according to the earlier review on literature, the type of VAR employed might exert influence on the reported results' variation. Regarding the present selection of studies, it has been registered a very extensive array of variations to the VAR methodology (see Table 3-2); these often introduce either specific features relative to new approaches or combine several modalities at one. In order to shorten the list, the choice was to summarize the types of VAR available by grouping them according to a prominent feature. Two groups, despite of discriminated at first, are characterize by using Bayesian inference methods in its process – TVP-VAR and Bayesian VAR. *Switching* models present an intern mechanism that enables to distinguish between ZIRP and normal regimes. The rest of the VAR model types are grouped in one category that includes VEC models. This choice of categories accounted for the limitations set by the scarce number of observations for some of the typologies, e.g., if we consider the FAVAR methodology alone, it would account for three entries in the database. Following the same reasoning, TVP and Switching VARs were grouped in one category with Bayesian SVAR (one observation).

Table 3-2: Types of VAR Methodology found in the Literature by Categories

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

Variable(s) that measure Output and *Other Variables used in the Regression* – The next set of data, presented in Table 3-3, is formed by all the variables used in each model described in the literature selection, from which the estimated impulse responses were produced. The first set (A), is composed of variables that vary greatly from study to study. To synthesize the collected information these were discriminated by their nature and fit into sub-categories[11], enabling a more parsimonious comparison. Moreover, the collection of this group of variables respected a two stage-process: first, one asks if the variable is used or not[12]; second, if it is used, it is assigned to a sub-category. An important note must be added in regards to the *Monetary Variable* category. In this context, monetary base is a broad sub-category that includes not only the estimates that used the monetary base but others that used one of its sub-components; in this case, they are either a form of estimation of the BoJ's Outstanding Current Account Balance (CAB) or Reserve Balance or Ratio[13]. In the same way, the money stock not only accounts for the estimates that did use a variation of the money stock but also 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). Variables that showed close resemble between themselves were synthesized into broader concepts (second subset, B, presented in Table 3-3)[14]; the third subset of variables (C) comprises those that did not require the need to be differentiated into subsets, since they do not belong to any specific category.

Table 3-3: List of Variables that Compose the Models Reported in the Literature

A) Categories of Variables and their Respective Sub-categories:

Variable(s) that measures the output	Monetary Variable	Secondary Monetary Variable
- GDP	- Monetary Base	- M2
- GDP growth and Output gap	- Money Stock	- M3

[11] See Table 9-2, Annex I for the full list of variables registered from the literature selection and subsequent designated category or sub-category (when applicable).
[12] A dummy is used: = 1 if used; = 0 if not. For a more comprehensive view on the matter, see Table 9-2 in Annex I.
[13] Reserves in this context are a sub-component of the BoJ's Current Account Balance, usually referring to the accounts that private banks hold on BoJ. A subsequent partition of this sub-component ,which is explored in the literature, is the amount of those accounts that is required to be held by law and those that are not (excess reserves).
[14] 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.

- Industrial Output
- Unemployment Rate

Price level (or Proxy)	Interest Rate of Reference	Exchange Rate
- CPI - Interest Rate - Core CPI Inflation Gap - GDP Deflator	- Call Rate - 3-month interest rate - Repo Rate	- Nominal Yen/Dollar Spot Rate - Nominal Effective Exchange Rate - Real Effective Exchange Rate - Trade Weighted Real Effective Foreign Exchange Rate

Spread	Bond Yield
- Difference between the 5-year JGB yield and the Call Rate - Difference between the 10-year JGB yield and the Call Rate	- 10-year JGB Yield - JGB Yields

B) Synthesized variables (= 1 if used; = 0 if not)

- Stock Prices (or Stock Price Index)
- Bank of Japan Stock Purchases
- Bank of Japan Bond Purchases

C) Other Variables (= 1 if used; = 0 if not)

- 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
- 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

- 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)
- 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

It is worth to mention that, in the first group, the *Secondary Monetary Variable*, only registers the variables M2 and M3, which were also included in models that already had a first variable of the same kind. In opposition to the variables in *Monetary Variable*, these were not regarded as monetary policy tools. The only exogenous variable found among the studies was *CPI inflation of energy and food*[15]. In the second group of variables (B), *Indirect Observance of Bank of Japan Monetary Policy*, accounts for synthetic variables build by researchers, which intent to indirectly observe the BoJ's policy stance over time.

IRF Window in Months – based on the temporal length of the estimation, we distinguish from the focus on short-term – until 24 months –, medium-term – between 24 and 48 months –, and long-term – more than 48 months –, (excludes TVP-VAR based estimates).

Type of Shock (1) – the initial goal was to qualify the disturbance in the monetary policy variable in three ways: which actual variable within the author's model was hit by the disturbance; the technique used to produce the disturbance; and its magnitude. Due to the fact that a significant portion of the shocks reported is not the usual 1% or one standard deviation (SD) increase in a given MP variable; and because authors often test several different policy tools, e.g., call rate and or a money stock proxy, for a period

[15] By exogenous variables, we are referring to those variables whose values are found outside the VAR system. The only study to use it was Dekle and Hamada (2015). By setting pre-determined values, the authors intend those variables to affect the system of equations, arguing that, CPI inflation of energy and food should affect the inflation rate with certainty.

under money supply targeting; the task was at first to characterize the shocks, when possible, in the previously designated terms, and then aggregate them by the affected variable (see Table 3-4). Shocks are commonly applied to the call rate, although, substitutes were used mainly because whenever the time of analysis comprehended ZIRP periods, some authors opted by other short-term interest rates [16]. It is also important to make notice that besides *Other Types of Shock*, it is implied that the shock is regarded as positive or a percentage increase[17]. 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*. These shocks, according to the variable they hit, can be thought as three different stages along the same transmission line, being the common goal to increase the money circulating in the economy. Shocks to BoJ's Current Account Balance or Average Outstanding Account Balance (AOAB) are a direct reflection of QE operations that are thought to affect the banking system reserves and then the money stock, before it hits output; shocks to bank reserves (or reserve rates) are an implied consequence of QE operations, and it is also expected that they'll 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 tried to relate the impact of financial operations directly related with the Large Scale Asset Purchase programme, by assessing the effect of government bonds purchases in the output. Michelis and Iacoviello (2016) were the only authors to resort to an approach that tried to quantify the required level of inflation inflicted by the BoJ, in order to promote output increase. Shocks that were not able to be categorized with the previously mentioned elements, and that do not fit in any of the latter described types of shock, were registered as "Other Types of Shock"[18]. In order to allow this category to be econometrically modeled it was required to short-down the list of possible types of shock; therefore, we re-organized it in broader categories. The criteria used to group these types of shock follows the one used to group the *Monetary Variable* category. The Shock to a Short-term Interest Rate of Reference (SSTIRR) includes the

[16] Usually the 3-month rate or the repo rate. Although not considered here, Nakajima (2011a, 2011b), employs a *shock to the medium-term interest rate gap,* which translates into a shock in the log-difference between the 5-year JGB yield series and the trend, computed using the HP-filtering.

[17] If some authors used the traditional one percent increase in the MP tool, others used proportional percentage increases to actual money supply targets, such as the current account balance, average outstanding account balance or reserves and reserves rate.

[18] Due to their complexity and heterogeneity of approach, this category includes the shocks reported and described by authors as shocks identified by the restrictions on impulse responses, because they are not easily comparable with other methods.

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. Despite of the loss of detail, the types of shock that do not fit any of the previous categories had to be included under the Other Type of Shocks (OTS) category.

Table 3-4: Type of Shocks found in the Literature Selection and how they were grouped

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) – often, the impulse response functions found in the selected literature are accompanied with intervals of confidence in order to assert the statistical significance of the disturbed variable in a given period segment. These intervals vary in terms of their process of attainment and width. There are three approaches that are used to define these intervals; although two of the definitions of the intervals found in the literature selection are equivalent: many studies define the intervals of confidence in terms of percentage – 95, 90 and 68%. On the other hand, another portion of the studies define the intervals in terms of standard deviations; the most common, ±one- and ±two-standard deviation confidence intervals, are roughly equivalent to 68 and 95 percent confidence intervals (respectively), assuming normal distribution. There is another form of confidence intervals, which uses the notion of percentiles. In the studies that use this type of intervals there have been registered two sets – the 10^{th} and 90^{th} percentile confidence intervals and the 16^{th} and 84^{th} percentile confidence intervals (PCI) – which are commonly produced by means of bootstrap techniques. Moreover on the aspect of equivalences, Primiceri (2005) is often cited to refer that under the assumption of normal distribution, the 16^{th} and 84^{th} confidence intervals correspond to a ±one-standard deviation confidence interval. For the sake of comparability, the equivalences were made so that the database registers 68, 90 and 95% intervals. Estimates with 10^{th} and 90^{th} 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 *a priori*, but rather discriminated, allowing for an eventual comparison between a broader and a narrower sample.

Are the Output and Monetary Policy Variable in Levels or in First Differences? – for the current analysis is pertinent to verify if the output and monetary policy variable to which the shock is applied are found in levels or in first differences, in order to justify the occurrence of explosive behaviors in impulse responses.

The Date of the Beginning of the Shock (if applicable)[19] – this permits to identify the studies in which there is impulse response functions that are set to affect a concrete period in time; making possible to compare, for the same period of time, the factual data with the estimated results through the model's simulation. One condition to register the entries with this specification was that the period which is affected by the disturbance should coincide, with at least one of the QE programmes, e.g., a shock set to affect the data starting at 2002:Q1 and onwards. This variable also serves to frame TVP-VAR impulse responses, which are usually reported in a tridimensional perspective. In these cases, for each time unit, e.g. month or quarter, is computed an impulse response; then a few values, e.g. 1st quarter, 4th quarter, 8th quarter, etc., are taken from each one of those impulse responses, to build single lines that allow to observe the behavior of all impulse responses, during the full period of analysis, after three months, one year, two years, etc., of the initial shock. Because we are only interested in the impulse responses that were computed during QE periods, we only considered, whenever available, the periods of 2002-06 (first QE programme[20]), 2010-11 (CME) and 2014-(…) (QQE). To set an example, Kimura and Nakajima (2016), analyze the following period, 1981:Q2 to 2012:Q3, for which they report a single line that depict the behavior of 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 (later registered in the magnitude category). 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 and 2005:Q4, 2009:Q4 and 2010:Q4[21].

3.2.4 Estimates

The following set of information is composed by elements taken from the observation of the impulse responses, which are displayed graphically in the selected literature. These elements attempt to summarize the most visible aspects of important consideration, to understand how, and if, monetary policy shocks have been affecting

[19] This variable, is preceded by a dummy variable that distinguishes the entries that actually use this specification from those that do not (see Table 9-2, Annex I).
[20] Because the first QE only starts at March 2001 and we are registering the shocks at the beginning of the year, we do not consider this year and start to register at 2002.
[21] If the data is presented in quarters, the beginning of a year is equivalent to the value of the 4th quarter of the previous year, e.g. 2002 equivalent to 2001:Q4.

output. These elements were registered only for impulse response functions that were previously acknowledged and registered as statistically significant at some point during their length.

Statistical Validity of the Impulse Response based on the Granger Causality (through the observation of the confidence intervals) – this “yes/no” variable has the purpose of assessing if a given impulse response is statistically significant during a time segment within the window of observation.

Signal of the Shock's Impact in the Output Variable – The intention here was to capture the overall effect of the monetary policy shock in the 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. Although we attempted to collect the estimates with due caution, the fact that some hump-shaped impulse responses have shown both positive and negative behavior, during its length of significance, made it more difficult to (visually) assess accurately the net impact in output. Moreover, to code the results of this variable was necessary to disentangle one more problem caused by the existence of different monetary policy shocks. The problem resides mainly in the fact that depending on the nature of the policy tool, the reasoning behind the inference also changes. Shocks based on money stock tools are aligned with the expansionary policy of quantitative easing, meaning that authors apply an increase in this variable and assess its impact on output. This is the basic reasoning of inference of these types of shocks. For accommodative policy tools such as the use of the call rate, authors approached the reasoning of inference in another way; by definition whenever the BoJ intends to engage in an accommodative policy, it will reduce by a percentage the yield of a short-term interest rate of reference. Authors, when simulating the impact of a shock to the short-term interest rate, even during the QE period, invert the inference process; they apply an increase in the interest rate and register its impact on output (contractionary shock), but when inferring about that same impact, they are trying to prove the opposite, e.g., if an increase of one percent in the call rate diminishes the output by a certain amount, the opposite would be also true (a reducing of one percent in the call rate would similarly increase the output). Therefore, a first task prior to coding was to identify and categorize the shocks by the nature of their reasoning. One category is designated as QE shocks – shocks that assess directly the impact of an increase of a QE policy tool on output; another category is the Non-QE shocks (contractionary and accommodative)[22], which is composed of interest rates that served as MP tools; and *Others*, a category used to identify shocks that do not fit in any of the first two categories but follow the same direct inference of QE shocks. Finally, because the main purpose of registering the sign of the shock was to assess the overall impact of a given monetary policy tool during the QE period, regardless of their type, we registered as "1", the significantly positive QE shocks (see Table 3-5 below); significantly positive accommodative Non-QE shocks; and also 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.

Table 3-5: The code of the overall effect of the monetary policy shock in output (Signal of the shock's impact in the output variable)

1	0	-1
Statistically significant: - Positive QE shock - Positive accommodative Non-QE shock - Negative contractionary Non-QE shock	Non-significant statistical estimates	Statistically significant: - Negative QE shock - Negative Accommodative Non-QE shock - Positive contractionary Non-QE shocks - Null values

Accumulated Effect of the Shock's Impact on the Output Variable[23] – some studies in the literature selection present graphically impulse response functions that are accumulated, in opposition to the non-accumulated. This factor indicates that the response of the

[22] If most authors that analyzed Non-QE shocks preferred to infer the results of contractionary shocks (increase in the interest rate) by inverting the results, other authors did apply accommodative Non-QE shocks (decrease of the interest rate). The inference on these last shocks is done strait forward, therefore, their sign is registered in the same fashion of a QE-shock. This distinction between QE, Non-QE and Others, appears in Table 9-2 (Annex I) under the category *Type of Shock (2).*
[23] This variable is also a dummy: = *1* if the effect is accumulated; = *0* if not.

variable observed, in this case the economic activity proxy or output, fades away not by converging to zero, but when the variable's fluctuation is decreasing over time. As stressed by Pitzel and Uusküla (2007), this practice although useful to dismiss the already stated accuracy concerns in assessing the net impact of the shock, it was not common to find in the literature.

Persistence of the Shock's Impact in the Output Variable (in Months) – the intention here was to register the temporal length for which the impulse response of the output variable is statistically significant. Due to the fact that the extraction of such information from the solo observation of the graphics which depict these impulse responses is an invitation to inaccurate sampling, the data extracted was registered under a cumulative sequence of two months at a time, e.g. at least two months; two to four months; four to six months; etc. This approach aims to reduce the level of inaccuracy but still, does not assure all the precision[24].

Magnitude: Value of the Shock's Impact in the Output Variable (in percentage intervals) – the variation of output has been registered for publication bias screening assessment purposes (Section 5) at 3rd, 12th, 24th, 36th and 48th month after the shock's hit[25]; and has been registered at its maximum value when the impulse function as a statistically significant period[26]. 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, see Table 3-6, are disposed as a cumulative sequence of 0.05%. In this way we hope that were 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%.

[24] Whenever the entries are not statistically valid because the Granger Causality is not verified earlier on, then persistence is registered as "NS" – Non-significant.
[25] These values were registered regardless of statistical validity. The moment zero has been registered separately as well, when available (contemporaneous shock).
[26] Marked as "NS" – Non-significant, if there are no statistically significant periods.

Table 3-6: IRFs Magnitude – Intervals of values

Nº	Interval	Nº	Interval	Nº	Interval
0	**[0]**	**11**	**]0.5 ; 0.55]**	22	]1.05 ; 1.1]
1	**]0 ; 0.05]**	12	]0.6 ; 0.65]	23	]1.1 ; 1.15]
2	**]0.05 ; 0.1]**	13	]0.55 ; 0.6]	24	]1.15 ; 1.2]
3	**]0.1 ; 0.15]**	**14**	**]0.65 ; 0.7]**	**25**	**]1.2 ; 1.25]**
4	**]0.15 ; 0.2]**	**15**	**]0.7 ; 0.75]**	**26**	**]1.25 ; 1.3]**
5	**]0.2 ; 0.25]**	**16**	**]0.75 ; 0.8]**	(…)	(…)
6	**]0.25 ; 0.3]**	**17**	**]0.8 ; 0.85]**	**34**	**]1.65 ; 1.7]**
7	**]0.3 ; 0.35]**	18	]0.85 ; 0.9]	**35**	**]1.7 ; 1.75]**
8	**]0.35 ; 0.4]**	19	]0.9 ; 0.95]	**36**	**]1.75 ; 1.8]**
9	**]0.4 ; 0.45]**	20	]0.95 ; 1]	(…)	(…)
10	**]0.45 ; 0.5]**	**21**	**]1 ; 1.05]**	**80**	**]3.95 ; 4]**

1) Intervals actually registered, for statistically significant, estimates marked in bold.
2) Values in absolute terms.

Transmission channels thought to affect output – we wanted to relate the conclusions withdrew from the observation of the (statistically significant) impulse responses depicted in the literature selection, with the transmission channels responsible for such results. As it will be noticed further on, many studies did not go beyond the task of proving the existence of a general causal relation between the BoJ's stance and the economic activity in Japan; notwithstanding, whenever the task evolved the refinement of the earlier premise, it became important to acknowledge the different approaches found in the study selection, regarding transmission channels; which may be explicit in the model – through the inclusion of a variable which embodies that very same function of transmission; or can be implicit – if the authors justify its presence with the support of economic theory and other empirical evidence. Transmission channels vary in their nature; often, these are broadly categorized as either a form of expectation or financial mechanism. Moreover, it was found in the literature selection that wider categories of channels were sometimes decomposed into sub-channels; in other cases, a channel could be isolated or refined into a more specific mechanism. The following list comprises identifies a transmission the transmission channels or effects identified in the literature selection. Because authors may point out more than one channel, these were:

Transmission channel undefined – this designation characterizes empirical results which do not define the type of specific transmission mechanisms, but rather assume the existence of a causal relation between a MP variable and output.

Interest rate channel – some authors regard the role of interest rates not as a target or a tool but as a mechanism capable of influence economic activity. For Schenkelberg and Watzka (2013) it is clear that the lowering of the real interest rate as a result of a successful forward guidance policy may influence positively the aggregate demand. Nakajima *et al.* (2011) state that an increase in the monetary base may indicate a shock in money demand if it follows the rise of short-term interest rates. For other authors, such as Nakajima (2011a)[27], the inclusion of a medium-term interest rate in a model that accounts for the ZIRP regime led this author to find evidence of an underlying policy commitment effect.

Forward Guidance – this channel is defined by the possibility of the BoJ to affect private agents' decisions regarding economic activity, via expectations. This channel aggregates two notions – policy duration/commitment and signaling. The first notion, for the Japanese case, means that the BoJ informs private agents of a plan (policy framework) to achieve actively a certain goal, thus influencing the present and future actions of those agents. Signaling in its turn is a notion subject to slightly different interpretations. For Ueda (2013), forward guidance is synonym of signaling effect in context of a large-scale asset purchase operation, in which the BoJ communicates the intention of continuing in the near future with such operation, transmitting commitment (guidance) to the agents in the economy. The signaling effect manifests itself when agents within the economy reduce their expectations on the future path of short-term interest rates; moreover this effect is usually regarded by authors, as being subdued to BoJ's specific actions. Honda (2014) and Ugai (2015) argue that the signaling effect may arise from the incessant increase in the monetary base due to BoJ's balance account sheet rebalancing or through a large-scale asset purchase programme. In addition, Shirai (2014) states that a firm public resolution by the BoJ to achieve a certain goal, may function as a signaling effect towards private agents.

[27] This study was not included in the literature selection from which the database was build.

What matters in the present case is to point out that it has been only considered forward guidance as an underlying channel of a given impulse function, whenever the study explicitly refers this channel has being present in form of policy duration commitment or/and signaling.

Inflation channel – inflation is thought to affect the behavior of private agents when predictions meet expectations. In the specific case of Japan, the struggle against stagnation is the main goal; therefore, if BoJ’s policy measures are followed by a better economic scenario, complying with the objective of higher sustained inflation, then a stronger forward guidance may be produced.

Effect of inflation targeting – closely related with the previous channel, some authors approached the question of what should be the BoJ´s inflation target level necessary to produce a relevant increase in output.

Stock price channel – this channel may be referred when an impulse function applied to a model that includes a stock price variable, corroborates a statistically significant effect of a monetary policy shock in the output. Notwithstanding, there are studies that decompose the stock price channel into more specific transmission mechanisms. For the case of Japan there are at least two sub-channels that are explicitly mentioned in the literature selection, despite of not being solely corroborated by the use of VAR methodology[28]. These sub-channels are[29]:

- *Wealth effect* – the increase in the value of stock prices may lead the owners of such asset to perceive themselves wealthier, which in its turn may lead these agents to engage more intensively into spending. An increase in demand can this way, exert a positive pressure in the output.
- *Tobin’s Q effect* – this effect is one that leads firms to invest more because of an increase in their intrinsic value via stock prices. Hence, this effect applies to firms that increase their total market value in relation to their total asset value, as

[28] The authors of such studies do not find evidence on these specific transmission mechanisms through the computation of impulse responses, but acknowledge the presence of an active stock price channel. The argument that this channel may be decomposed into more specific mechanisms is generally supported with other theoretical and empirical evidence within the same studies.

[29] Although there are other sub-channels associated to the stock price channel, the ones presented were the only suggested to work for the case of Japan.

a consequence of an increase in the stock prices of a successful stock issuance operation.

Portfolio re-balancing effect – this channel is used to refer an effect that leads firms to transact more intensely in the financial markets. One of the objectives of the quantitative easing programmes imposed over time by the BoJ had a specific intention of helping to mitigate the disruption of the financial intermediaries' role in the system. In the same line of thought, a more active portfolio re-balancing within the private sector, implicates other positive side-effects as portfolio risk smoothing, increase in financial returns, and higher capacity to deal with non-performing loans. Other channels found in the literature selection that may fall under the portfolio re-balancing definition:

- Increase in excess reserves – according to Honda (2014), an increase in excess reserves of the lending sector as part of a general injection in the monetary base, foments the desire in the private sector to demand assets that are imperfect substitutes for money. Such demand would eventually increase asset prices and force private agents to change their portfolio composition, which as a final act may create spillovers onto the real economy.
- *Bank Balance Sheets* – bank balance sheets that are affected by increases in the excess reserves, during the QE periods, lead banks to engage in more lending activity and bond purchasing.
- *Asset Purchase Program* – in this specific case, authors intended to quantify the effect of specific elements of the QE programme in the economic activity. The *Asset Purchase Program* may be decomposed into large-scale asset purchase operations, which have an underlying effect on the composition of private agents' portfolios and are associated with the growth of the monetary base.

Credit-easing channel – this channel, although often cited in the literature, it is not usually treated as a phenomena capable of being explicitly isolated through VAR methodology, but is rather treated as an implicit channel that may be associated with other transmission mechanisms. The reason behind this is justified theoretically, by stating that when the central bank engages in quantitative easing it expands its own balance sheet, the asset side, by purchasing several types of financial assets; the credit easing takes place when some underlying mechanisms within the financial market respond to the expansion of the central bank's balance sheet, providing a better credit environment for firms and households. Those mentioned underlying mechanisms may have multiple sources; for Ugai (2015), credit easing is associated with the smoothing effect the BoJ provides to unbalanced markets, where the expansion of its balance sheet is done to the expense of low liquidity asset purchases. This type of asset purchase is also said to reduce the risk of the sellers' asset portfolio, which can be seen as a portfolio re-balancing effect. Ueda (2013) adds that term-loans conceded by the BoJ to stanch the risk premium component of asset prices, may function as a credit-easing channel. For Kimura and Nakajima (2016) the credit-easing channel is working whenever the expansion of the BoJ's balance sheet lowers market long-term interest rates or spreads. Despite of what was previously said about the credit-easing channel, this one would be in fact, more implicitly found on the previous set of channels regarding the *portfolio re-balancing effect*. The rule of thumb is: if there is an increase in the monetary base or an increase in the excess reserves of banks at the expense of an expansion of the BoJ's balance sheet, than there might be underlying credit easing mechanisms at work. Albeit this reasoning is seen as consensual, it was not widely explored neither mentioned in the literature selection.

4. Descriptive Statistics

The descriptive statistics provide a first characterization of the data collected from the literature, and its resume is presented in the Table 9-2, Annex I. Regarding the authorial aspects of the sample, and specifically the year of publication, it is noticeable that there is a great concentration of collected estimates (entries per study) in the right side of the central value – the mean falls approximately in 2012 (see fig. 9-a, Annex 1); despite of studies ranging in terms of year of publication from 2006 to 2016, the midpoints of the studies' data timeframe do not go, on average, beyond 2005[30]. Considering the full sample size – 104 observations – a relevant portion of the research debate is still made outside the publishing sphere[31] – 54.81% – of the sample is composed of working

[30] Due to methodological differences, depending on the type of models, the midpoint of the studies' timeframe can differ greatly; the registered average midpoint, were (approximately): for VAR-VEC and BVAR – 2005, TVP-VAR – 1999, and MSVAR – 1995.

[31] From the 30 studies collected, 15 are working papers or mimeos, 10 are published in general journals; and 5 are published in monetary journals.

papers and *mimeos*; the portion of the sample that has been published in general journals, is 38.46%, and in monetary journals 6.73%. It is worth mentioning that a sizeable portion of the observations, 20.19%, were made available by studies from which at least one author is (or was) directly associated to the Bank of Japan.

The data used in these studies, do share some similarities; more specifically, 60.58% of the data used to produce the impulse responses comprehends the years of the first QE programme (2001-2006). This is rather understandable having in mind that at least 50% of the studies were published until 2013, but it also stresses the fact that less attention has been given to the understanding of the individual impact of the later QE programmes. The favored periodicity has been monthly data – 53.85% –, followed closely by quarterly data – 44.23%. Daily data has been used only by Matsuki *et al.* (2015), which used data on domestic daily power consumption has the output proxy. We also registered that the number of observations that constitute the timeframes of analysis, varies as much as from 20 to 336 observations; around 44% of the estimates registered were produced using samples with over 100 observations; 43% have between 50 and 100 observations; and the remaining 12.50%, below 50 observations.

Relatively to the VAR approach used in each study, it was already stated that there was a loss of accuracy by shorting down the list of the many VAR model variations found on the literature, to a degree that enables that information to become sufficiently parsimonious to be modeled within the meta-analysis context[32]. Notwithstanding, we may add that within the TVPVAR-BVAR-Switching group, the TVP approach, considering all its variations, has more relevance (27.88%) than the Switching VAR group (9.61%)[33] or the Bayesian VAR (one observation). Similarly, entries in the database marked as the basic VAR, account for 34.62% of the wider VAR-VEC group, which accounts for 62.50% of the whole sample.

Looking at the variables contained within those models, the most used variables, besides output, are, in descending order of relevance: Monetary variable (89.42%), Price Level (85.58%), and Interest Rate of Reference (53.85%); followed by the Exchange Rate (26.92%), Stock Prices (24.04%), and Bond Yield (18.27%). No surprises arise from these latter results, even though 33 different (categories of) variables were identified throughout the literature selection (besides the variable that

[32] For a complete account of the percentage distribution of the variations of the VAR model, see Table 9-4 in annex I.

[33] Were considered switching VAR models: MSVAR, MS-FAVAR and Regime Switching SVAR.

measures output). The unmentioned 27 categories of variables are not of standard use, thus possessing small relevance in the whole sample. Moreover, the number of variables in a model may vary between the minimum of 3 up to a maximum of 9. It may also be important to make notice that the type of shock that has been registered the most – Shock to the Monetary Base (52.88%) – actually owns its weight to the Shocks to the Outstanding Current Account Balance, which account for 48.08% of the total. The Shock to the Money Stock excluding the shocks to the JGB's held by the BoJ, account for 24.04%; and expectably, the Shock to the Short-term Interest Rate of Reference (12.50%) are mainly a reflection of the employment of shocks to the call rate (10.58%).

At last we make a prior analysis of the information gathered for the Estimates Section (3.2.4) of the database, which correspond to the characterization of the Monetary Policy shocks to the (Japanese) output in terms of signal (overall effect), magnitude, persistence, and what transmission channels may have been associated with those impacts. In terms of the overall effect of the shock, it is noticeable in figure 4-a that a great portion of the estimates – 50% – suggest an overall significant positive effect of the BoJ's capability to promote the increase of output (at some undesignated extent); whereas an almost similar portion of estimates – 47.12% – did not find statistical proof to support that result. Furthermore, only a residual number of estimates – 2.88% – point that the BoJ policies had a negative overall impact on output. To prevent the detraction from what has been reported in the literature selection, the (intervals of) magnitude and persistence of the behavior 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[34]. Concerning solely the density of magnitude of QE/Other shocks to the output (fig. 4-b), 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 small relevance, 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 positively the output, but in this specific case, with little to no relevance. Conversely, the same scenario may be traced for the Non-QE estimates (fig. 9-b, Annex I), despite of an inexpressive landscape provided by a small number of observations. In terms of

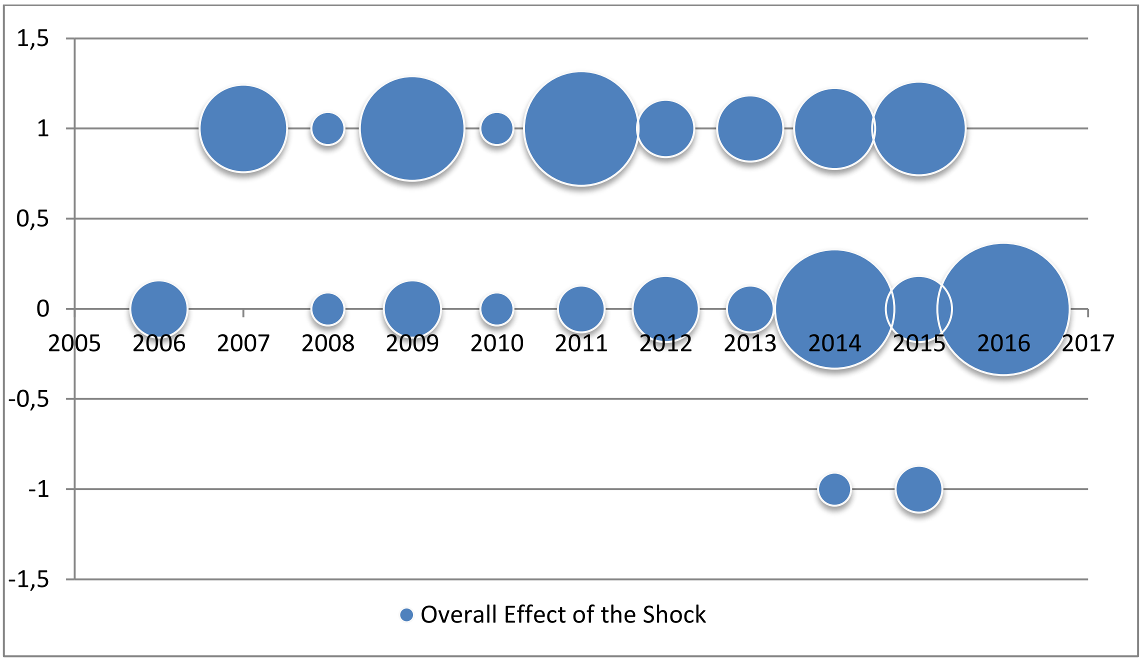
[34] According to these categories, out of 104 observations, 80.77% fall under QE; 12.5% under Non-QE; and 6.73% under Other Shocks.

persistence of the shocks**,** the number of observations is considerably lower – 33 for QE/Other estimates and 8 for Non-QE estimates; nevertheless, a majority of QE/Other estimates suggests (see fig. 4-c) that monetary policy shocks set by the BoJ in QE years are able to continuously affect Japanese output by periods that range from up to 2 months until up to 10 months. With only 7 observations, the persistence of Non-QE shocks does not allow us to withdraw any consistent and relevant information (fig. 9-c, Annex I), although it is possible to say that the distribution of these observations follows roughly a similar pattern to the QE scenario.

To complement the latter depiction of the information gathered around the elements that define the aforementioned shocks, Table 9-5 (Annex I) gives an account of the transmission channels that authors associate with the eventual effectiveness of those very same shocks. First off, 23 out of 30 studies report statistically significant results[35] and mention the existence of transmission channels at work. Also, no study points more than two transmission channels to justify the output response to the shock. A reference to an *Undefined Transmission Channel* is the most common association for any type of shock (12 out of 30 studies). Second, the shocks which appear associated with the use of QE tools comprise a wider number of possible transmission channels. The stock price channel, if considered along with its sub-channels – *Tobin's Q* and *Wealth Effect* – has been mentioned to be at work in 5 studies, while the *Portfolio Re-balancing Channel* has been mentioned explicitly or through its close related effects – *Increase Excess Reserves*, *Bank Balance Sheets,* and *Asset Purchase Program* or even through the *Credit Easing Channel* – in 7 studies. Even though highly mentioned in the literature as a typical working channel, *Forward Guidance* was not often associated with the effectiveness of shocks without being subdued to another channel; regardless, it appears explicitly mentioned to be at work at least in 2 studies. Studies that report statistically significant impulse responses by employing Non-QE shocks, only mention either an *Undefined Transmission Channel* (4 studies) or the *Real Interest Rate Channel* (1 study).

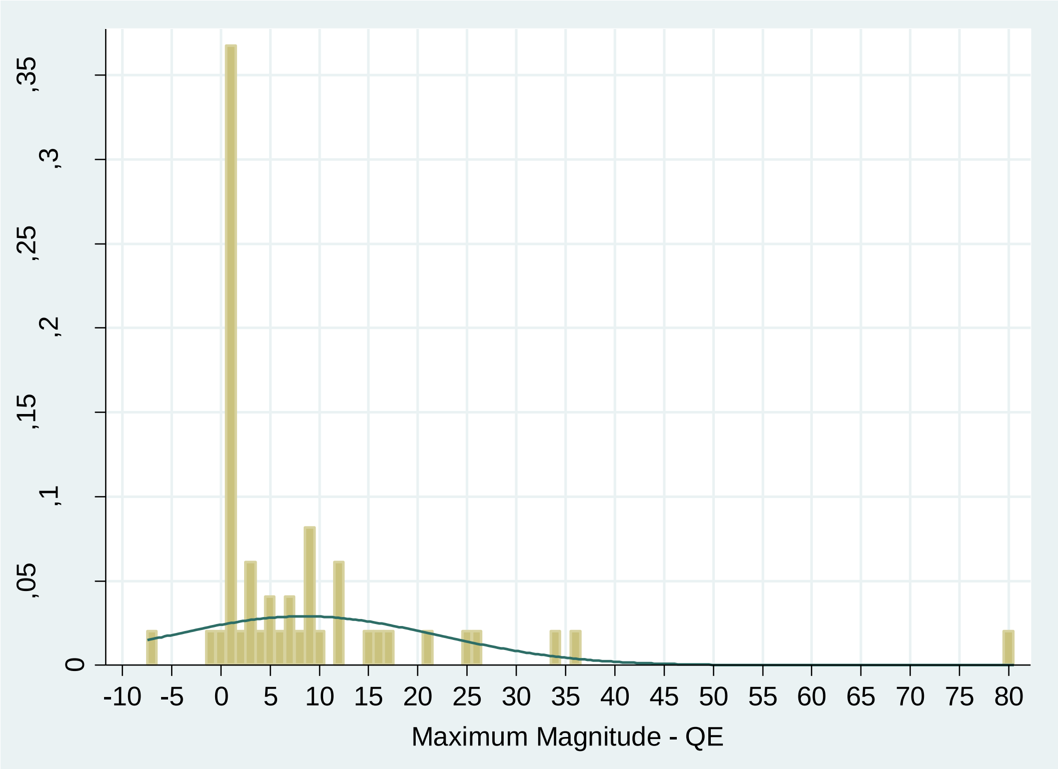
[35] Out of the results collected in the present study. Beyond those 23 studies, there were another four included in the selection that did report statistically significant results, but didn't do any particular association or reference to the concept of transmission channels.

Figure 4-a: Overall Effect of the Shock in the Output by Year of Publication/Release (104 Observations)



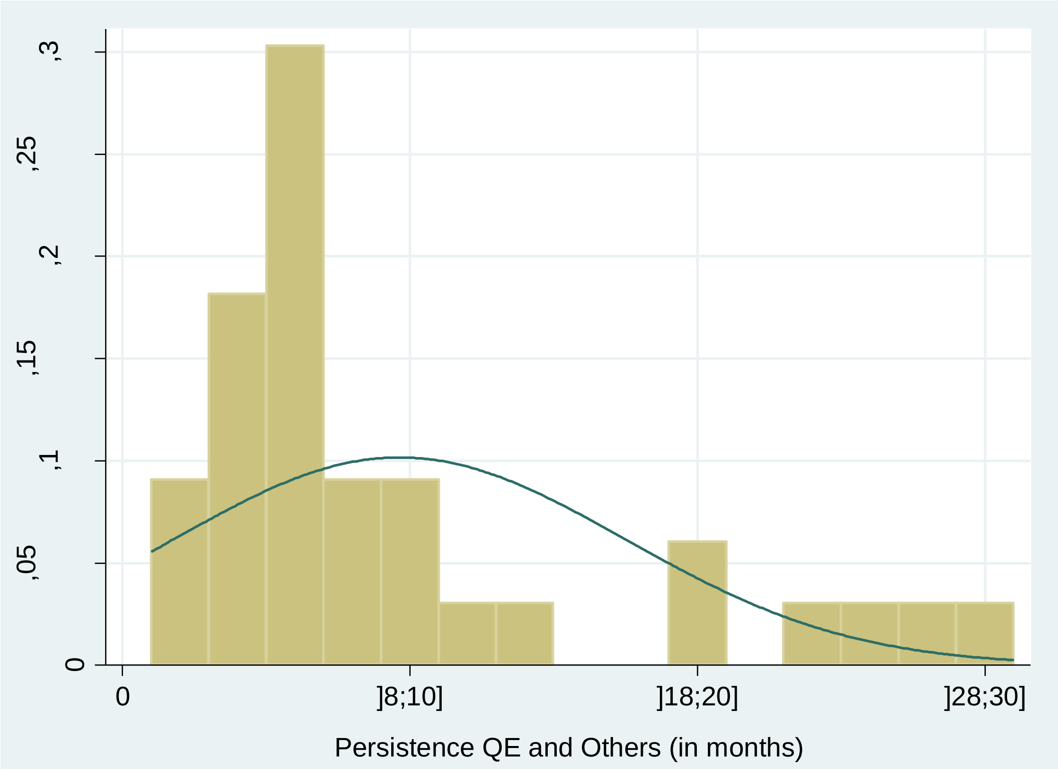
Note: The size of the marker indicates the density of estimates collected from the studies for each year. If "1", the effect is overall positive – 50%; "0" if the effect size is non-significant – 47.12%; "-1" if the effect is overall negative – 2.88%.

Figure 4-b: Density of Intervals of Maximum Magnitude of the Output Response to a QE Shock or Other Shock



1) The first ten (positive) intervals are read as follows (in percentage): 1 –]0 ; 0,05]; 2 –]0,05 ; 0,1]; 3 –]0,1 ; 0,15]; 4 –]0,15 ; 0,2]; 5 –]0,2 ; 0,25]; 6 –]0,25 ; 0,3]; 7 –]0,3 ; 0,35]; 8 –]0,35 ; 0,4]; 9 –]0,4 ; 0,45]; 10 –]0,45 ; 0,5].
2) Intervals correspond to the maximum value attained by the output response to a MP shock, during a statistically significant period (49 Obs.)

Figure 4-c: Persistence of the Shock's Impact in the Output Variable during a Statically Significant Period (in Months; 33 Obs.)



5. Publication Bias Screening

As expected, a meta-analysis exercise should include a section explaining 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 although, as it will be referred later, publication bias can be extended to a wider form of publication/releasing bias. 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 in one of the sides 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 is 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, in order to see their research published. Commonly, this behavior may manifest itself in a form of *cherry picking*, e.g., by inferring for the whole based on a very specific sub-sample or model. 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 either with the established theoretical or empirical history are less likely to be published. Picking up on this last idea, and even though not explored in the present study, Neves *et al.* (2012) makes an interesting assessment that bias regarding reporting results can go further as a consequence of a temporal pattern, designated as the economics’ research cycle[36]. According to this idea, researchers tend to replicate and report the results in conformity to a prior study, which has been considered a break-through in the given field. As posterior results may be tied to that major contributor, it is likely that other researchers tend to follow the methods undertaken by that seminal study until the next major contributor steps into a new cycle. This is important to alert that bias in studies may occur to a deeper level that may range from how the given subject is conceptualized to what methodology is used, and how these research aspects are justified. Lastly, when a subject as the one discussed throughout the present study is based on a sample of collected estimates, from which only 45% are taken from published studies, we must look upon the non-published studies, if the objective is to extensively characterize the literature about the subject. Moreover, it is not because a study was only released and not published by a third party that makes it impervious to publication/selection bias, even though, possibly in a lower degree (less conscious incentives perhaps).

One way to analytically screen for publication bias within a pool of estimates, henceforward also designated as effect sizes[37], is to relate the value of each estimate with a value that measures its estimation precision. The expected relationship, in the absence of any systematic distortion, is the higher the precision the less variation (with mean zero) around a “true effect”[38]. By “true effect” we mean an identifiable central value, from which effects-sizes may vary regardless the level of variation. The notion of true effect is important as a way to discern if a central value may be perceived as a proof that a given economic relationship, that has been studied, 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, due to the fact that authors may report

[36] Begg and Berlin (1989) make a similar reference but evoking the exploratory/confirmatory cycle.
[37] We are employing the term in its current sense, and not exploring its multiple interpretations. Read Kelley and Preacher (2012) for a comprehensive analysis on the concept of effect size.
[38] We borrow the term “true effect” from Begg and Berlin (1989) and Stanley (2001).

intentionally higher values to compensate less precision. The analysis that follows are embedded in this latter idea and comprise two types of tests: the Funnel Asymmetry Test (FAT) and the Precision Effect Test (PET) (Stanley 2005, 2008; Doucouliagos and Stanley, 2009).

5.1 Funnel Asymmetry Test

The first part of the analysis is more informal and is based on the analysis of scatter plots that put effect sizes against a measure of its statistical precision. Among the elements that characterize the impulse response functions – the overall effect (signal), persistence and magnitude, the latter is the most suitable effect size. The overall effect is not informative once it is defined as tridimensional – positive and statistically significant, non-significant, and negative and statistically significant, and along with persistence, it was bound to the existence of periods where the responses are statistically significant, in order for those to be determined. This is not desirable in the present analysis given the risk of type I bias being present, if statistical significance was "forced upon" the estimates[39]. 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 will be used as a rough approximation of the standard error (Stanley, 2005), whilst being aware of some possible caveats. First, there must be a prior assumption that the sample size is somewhat correlated with the measure of precision, which in this case is fairly acceptable once the number of observations is the denominator of the standard error, and according to statistical theory, the square of the first should increase proportionately as the latter diminishes. 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 to support that the chosen studies had pre-determined sample sizes, something unusual in this type of literature[40].

[39] As we shall see later on, we conduct tests for statistically significant effect sizes (those inside the confidence intervals) and compare the results with those, which do not regard statistical significance. When we refer that statistical significance may be "forced upon" estimates we are referring to the fact that some authors may choose wider confidence intervals if smaller ones invalidate their results.
[40] That is an approach more usual in experimental frameworks.

As mentioned earlier in Section 3, the drawback of using the number of observations is that these depend on the timeframes reported in the studies, which seldom report the number of observations for each particular model (and set of related impulse responses), hence, possibly distorting the analysis. To reduce the uncertainty regarding the sample sizes used, we exclude the entries within the TVP-BVAR-Switching category (except the one by Bayesian SVAR) since for these methodologies the reported timeframe of analysis may not match the sample size used to produce the models' parameters estimates. The closest study found, in terms of the methodology employed in our analysis – Rusnák *et al.* (2011) – uses a more direct standard error source by taking the magnitude length that goes from the value of the effect size to the correspondent confidence band that is closer to the horizontal axis, i.e., when the magnitude value is negative it is the upper band that must be considered, and vice-versa[41]. The procedure used by these authors seems to be more accurate than just using the number of observations, but their data besides being more homogeneous, is larger. As previously mentioned, it was very difficult to obtain the effect sizes by visual inspection and to determine the corresponding band values. Moreover, some confidence bands that were provided are not immediately convertible into a single measure, e.g. the 90 PCBs and the 10th/90th percentile CBs. It would also be necessary to convert the thresholds of all the confidence bands into the wider version found on the literature – 95 PCB (approximately two standard-deviations), otherwise, to convert wider confidence bands into smaller ones could determinate statistical insignificance (which would be an important aspect whenever statistical significance is relevant to the analysis).

When analyzing this type of scatter plots, 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 (a decreasing standard error for instance), the variation around a "true effect" size should diminish evenly. This relationship should translate, by visual inspection, a pyramidal shape of the scatter plot, or, as it was coined in the literature, an inverted funnel shape. The detection of publication bias is based on a plot feature that, besides having a funnel shape, must be assessed: type I bias may occur if the 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 about the characteristics of the effect sizes here depicted is that it would not be

[41] This implied that the estimates should be statistically significant.

prudent to analyze effect sizes, which 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 the output in a different order of magnitude once compared with QE tools, e.g. increase of the current account balances. They may present two different true effect sizes; hence, the best approach is to split the analysis into QE (and "Other" shocks) related effect sizes and Non-QE related effect sizes**.** In consonance with this idea, Stanley (2005) makes notice 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. Given that this is precisely the case here, all considerations in this section must be done with utmost discretion. All things considered, the analysis includes considerably shorter subsamples taken from the original 104 observations and the funnel plots displayed[42] concern only QE (and Other Shocks) sizes: the maximum magnitude when statistically significant and the magnitude values at the 1st quarter, 12th and 24th months horizons (see figures 5-a to 5-d). We also distinguish the effects presented at published papers from those of working papers and mimeos.

When observing the scatter plots [43], we are able to notice in all situations a high concentration around a small interval of magnitude values, to the exception of the Maximum Magnitude When Statistically Significant effect sizes[44], in which such pattern is less clear. This indicates that there might be a consensus in the literature around what can be designated as the true effect. In the Maximum Magnitude plot it is noticeable a funnel shape whether we are considering both published and non-published effect sizes together or separately. Also, we can notice that non-published effect sizes are more disperse than published ones, with special emphasis to the right half of the plot. In regards to 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 in positive territory. Twelfth Month and Twenty-fourth Month plots show similarities in the sense that in these two cases the funnel shape looks more evident if compared with the former

[42] Only 1 outlier has been removed from this analysis. 36th and 48th month effect sizes did not provided enough observations for robust estimations; therefore, they have been excluded from this analysis as well.
[43] Plot values (dots) may overlap.
[44] Here on after shortly designated as Maximum Magnitude effect sizes.

plots, being also noticeable that regardless of being published or unpublished, effect sizes are more disperse at the right of the main concentration of values. Also in both cases, with slightly more emphasis for the last plot – 24th Month –, there is more dispersion of non-published effect sizes in the right side of plot, 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 may indicate the presence of publication bias and, although the plots here presented do not constitute a definitive evidence of such, they point towards a biased scenario. Unfortunately, the lack of observations for Non-QE effect sizes left us almost nothing to work with. Since the solo observation of these specific plots might be a weakness in order to detect patterns that lead to infer on eventual publication bias, one ought to conduct a second form of analysis that may help to shed some light over this issue.

Figure 5-a: Effect-sizes collected from the Literature Selection – Maximum Magnitude

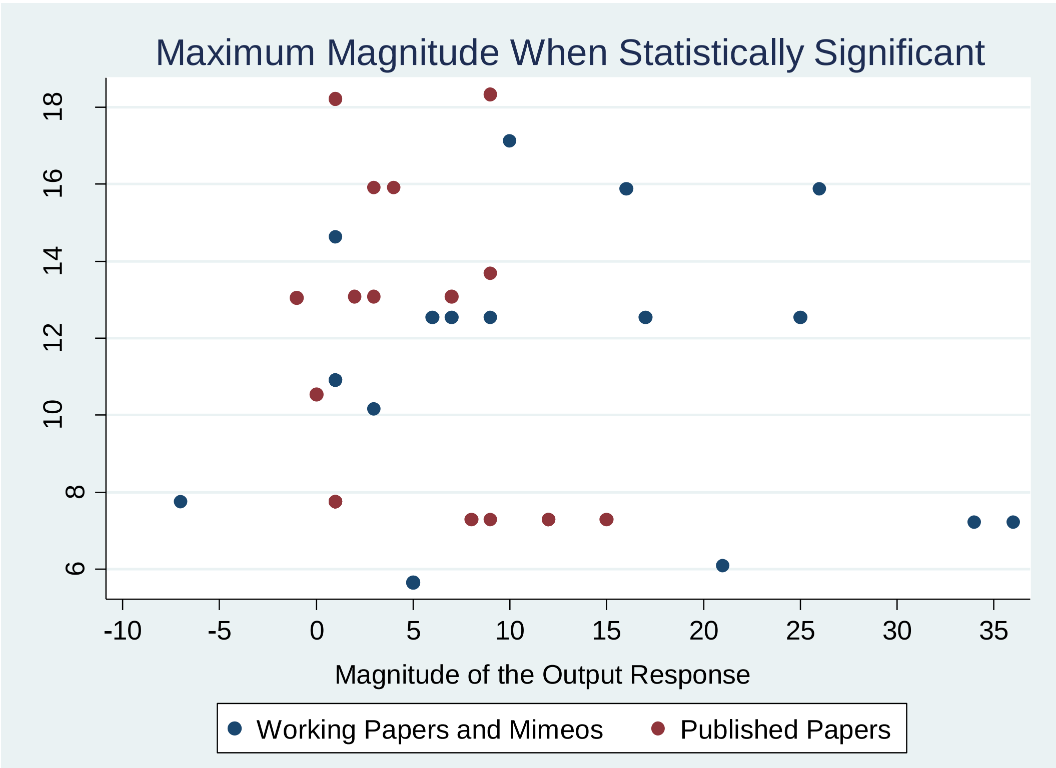


Figure 5-b: Effect-sizes collected from the Literature Selection – First Quarter

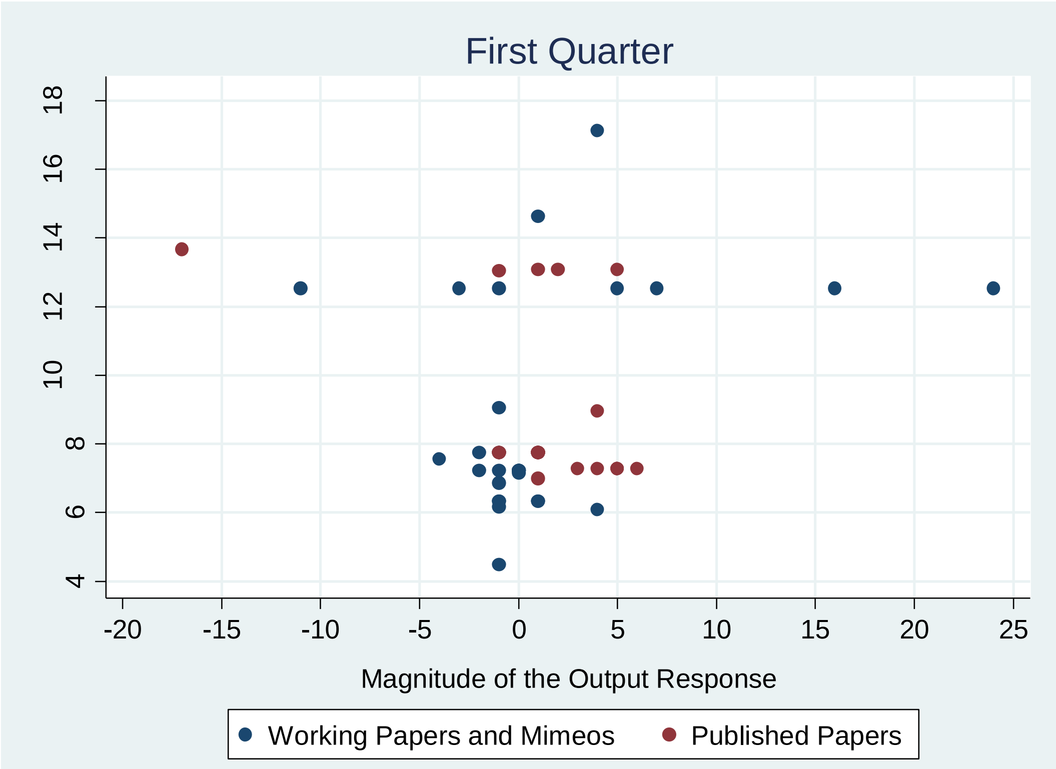


Figure 5-c: Effect-sizes collected from the Literature Selection – Twelfth Month

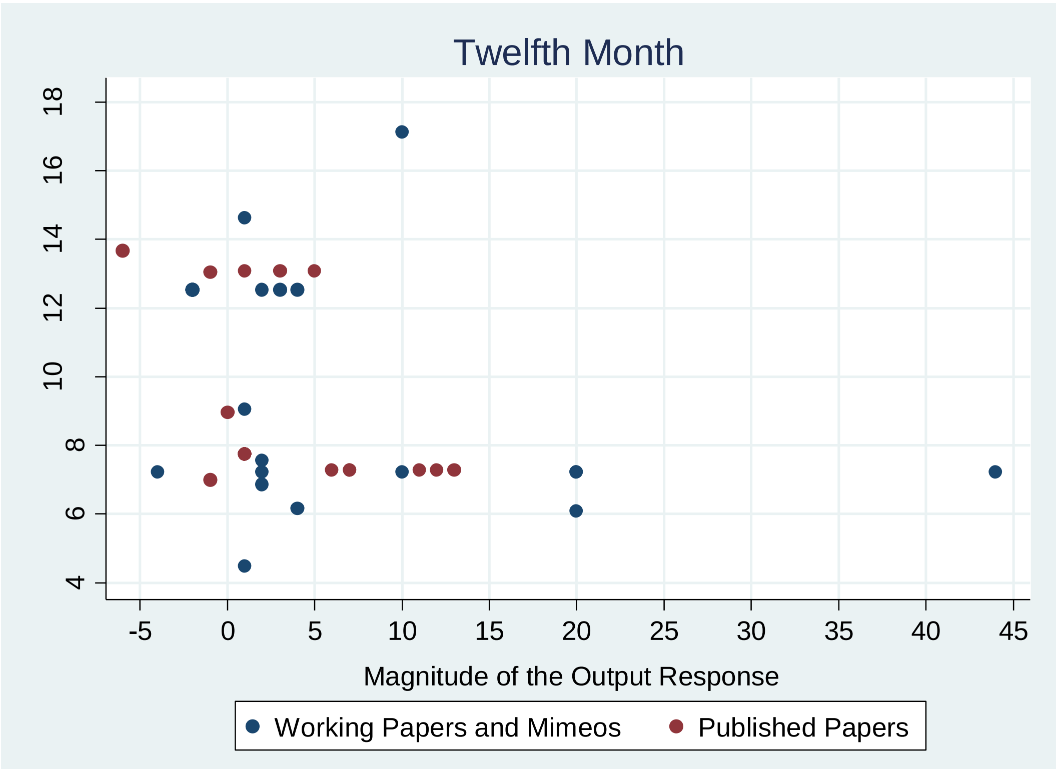
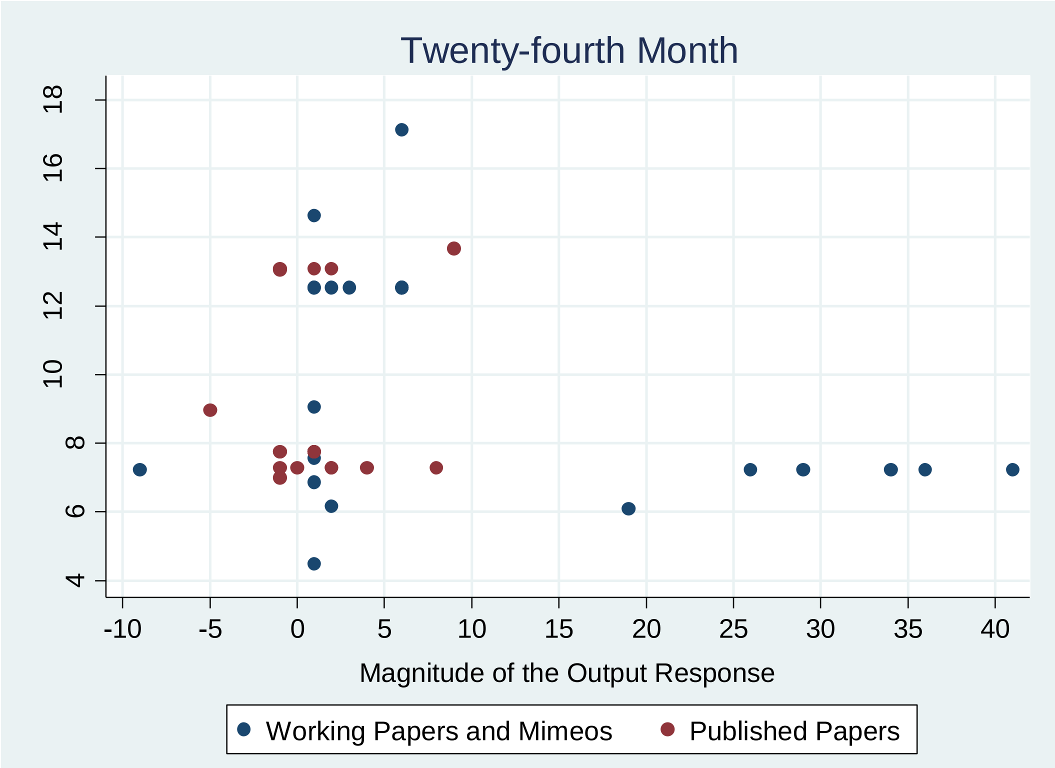


Figure 5-d: Effect-sizes collected from the Literature Selection – Twenty-fourth Month



5.2 Precision Effect Test

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

$$\widehat{\gamma_i} = \beta_0 + \beta_1 Se_i + \varepsilon_i, \qquad i = 1, \ldots, n. \qquad \varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2). \qquad (1)$$

In this first model, it is made explicit that the estimated size effect, $\widehat{\gamma_i}$, in our case the impulse response magnitude i, depends on its standard error, Se_i. Here, a true fixed value effect is given by β_0 and, whenever publication bias is absent, the correlation between the effect size and the standard error must tend to zero – β_1 should be statistically insignificant, otherwise, there could be bias. The error term ε_i, is normally distributed and independent across i. Consider the following model, where Se_i is equal to the inverse of $\sqrt{n_i}$:

$$t_i \equiv \frac{\widehat{\gamma_i}}{Se_i} = \beta_1 + \frac{\beta_0}{Se_i} + \varphi_i \Leftrightarrow \widehat{\gamma_i}\left(\sqrt{n_i}\right) = \beta_1 + \beta_0\left(\sqrt{n_i}\right) + \varphi_i, \qquad (2)$$

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

This specification (2), deals with the fact that there is a strong prior assumption that this specification is heteroscedastic: if we were using the real SEs as explanatory variables, then there would be an obvious correlation with the independent variable, since the latter is a sample estimate of the standard deviation of the former effect size, To circumvent this, (2) becomes the "weighted version" of the first specification, dividing (1) by their correspondent SE – in our case its proxy $(\sqrt{n_i})^{-1/2}$ – in which the right-hand side elements of the equation (excluding the error term) switch their interpretations: the true effect is given by the coefficient associated to β_0, and the impact of precision by the constant term, β_1. Also, the dependent variable, $\widehat{\gamma_i}(\sqrt{n_i})$, is now a proxy of the t-statistic, and the measure of precision is not inverted anymore (Stanley, 2005). Consider the following model:

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

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

The specification described in (3) is a generalization of the latter version that accounts for the possibility that the effect sizes are correlated at some extent within the same study[45], due to methodological and/or data similarities. This allows to acknowledge and quantify to what extent such correlation affects the model estimates. To this end, one can make use of the type of models[46] designated as multi-level mixed-effect (MLME) which can be seen as an extension of the simple linear regression models (OLS). In the present case, the MLME model is also of the simplest type, since we only wish to fit two levels – single observations (first) and a single stage of clusters based on grouping observations by study of origin (second). The observations are now discriminated as the i_{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 that is given by $\epsilon_{ij} + \alpha_j$, in which we assume that α_j, embodies a measure of within-cluster correlation. More so, that is said to be random because it is the variance at our specific cluster level and does not depend on the "fixed

[45] An alternate specification could be to cluster by author instead of by study.

[46] Also designated as hierarchical model once it belongs to the type of models that allows accounting for the correlation between groups of observations, in which clusters composed by the initial set of observations may be nested at a second (smaller) level of clusters. This procedure may be applied further into smaller cluster levels, depending on the adequacy to the data in hand.

part" of the model. Similarly, ϵ_{ij} is the given variance of the residuals at the base level – single observations (thus, the estimated variance of the overall error). The assumptions of this model assume that these "random" components are normally distributed and have their own constant variance and that these terms are independent and identically distributed, which in practice, and since ϵ_{ij} and α_i are the two components of the error in (2), they sum up to its total variance – $Var(\varphi_{ij}) = \delta_\alpha^2 + \delta_\epsilon^2$. Also, given that $Cov(\varphi_{ij}, \varphi_{gj}) = \delta_\epsilon^2$, $i \neq g$, the intra-class correlation[47] (ICC) between individual and cluster level is $\rho_\varphi = Cor(\varphi_{ij}, \varphi_{gj}) = \delta_\epsilon^2/(\delta_\alpha^2 + \delta_\epsilon^2)$ and assumes that the within-cluster error is equally correlated (Cameron and Miller, 2015). In this study, we compare the MLME models' results with the analogous versions of OLS with cluster-robust standard errors[48], which may be seen as an alternative approach to cluster modeling. Also, whenever there is no clear evidence that the MLME approach provides a better fit than simple OLS, we study this last specification.

5.3 Estimation Results

In the present case, due to a lack of observations, the number of clusters (studies) seems disproportionally high[49], even though the literature in which this exercise was based, seemed to pay little attention to this specific issue. In part, it is because MLME models are said to handle better the eventual lack of balance of clustered data; but also, because we are comparing the MLME to a clustered OLS, it is important to notice that few observations for a considerable amount of unbalanced clusters[50] are a source of bias in the parameter's standard errors for the latter type of model. Furthermore, the comparison of distinct models serves as a robustness exercise by checking if the findings are either in consonance or in disagreement. Table 5-1 gives us a resumed account of the main models estimations. First, we look at the results gathered from the effect sizes registered at its maximum value during a statistically significant period (Table 10-1, Annex II). We recall that the effect sizes are values taken from impulse

[47] Recall that in a 0 to 1 spectrum, 0 stands for no correlation within clusters (no variance to explain at cluster level) and 1 stands for maximum correlation (no variance at the individual level, i.e. the observations within clusters have the same value).

[48] Henceforth, designated as cluster OLS. This model specification is intended to prevent over-rejection of the parameter's statistical significance. If cluster OLS is more suitable to a given dataset and simple OLS is used instead, then the OLS standard error may be too small, thus producing high t-statistics and small confidence intervals (Cameron and Miller, 2015).

[49] For a battery of test results accompanied by data specs, see tables 1 to 4 in Annex II.

[50] For unbalanced clusters we are referring to the discrepancy of observations within each cluster.

response functions and these are statistically significant whenever the confidence intervals' thresholds are either positive, or both negative[51]. The exercise is performed for all studies and for published studies only – in the latter case we address the exercise as being a publication bias screening and in the former as a systemic bias screening. In order to evaluate the distributional assumptions, we firstly test whether there is evidence that the error term is normally distributed, which in this case we may not reject for published studies (Jarque-Bera test *p*-value over 10% for the Cluster OLS model and over 1% for the MLME version). Since we are more interested in the significance of the parameters than their estimated values, the non-normality of the errors prevents us from taking solid conclusions when that condition is not verified; notwithstanding, the results show a similar scenario when extending the analysis to the models that consider all studies. In both cases, MLME models are preferred over the simple OLS (likelihood-ratio tests below 1% suggest that we can reject the null hypothesis of a MLME equal to a simple linear regression); ICC values are over 0.5 (and below 1), reinforcing the idea that the cluster specification of the models is relevant. The estimation results reveal that if the cluster OLS version for published studies only is taken into account, the constant parameter is not statistically significant, which suggests that there is no publication or systemic bias. The non-constant parameter is not statistically significant as well which suggests that there is no true effect in any of the situations. The scenario here depicted for the "maximum magnitude" effect sizes mirrors the results gathered to the published 12th Month effect sizes (Table 10-3, Annex II), for both the MLME and Cluster OLS models. It reinforces the suggested evidence of no true effect size or bias. Regarding the models for 1st Quarter and 24th month effect-sizes (Table 10-2 and 10-4 respectively) there is still evidence of normally distributed errors at 1% when considering only published studies as well (MLME versions); the results from these models are in line with those previously found, although 1st Quarter model does not find evidence of a true effect at 10%. As an additional note on the 1st Quarter effect sizes, considering all studies suggests that the cluster specification does not seem suitable in this case (ICC is zero), i.e. there is no within-study correlation among observations. Without this latter specification, the model to use is a simple OLS which presents the same problem of non-normality of the errors.

[51] As previously justified earlier on this section, only the first set of estimations comprised in Table 10-1 (Annex II), has in consideration whether the effect sizes (IRFs values) are statistically significant or not.

Taking the wider view in regards to the later analysis, the first conclusion that we obtain is that there is a sense of accordance that the published literature on the previously mentioned subject does not provide consensus on what might be a true effect. This is a premise that must not be seen as definitive, once there is the chance that heterogeneity of methodologies and frameworks across studies may originate several true effects, i.e. preventing us to identify a clear central value when observing the funnel plots. Second, there is some caution that must be taken when applying this screening method to the VAR methodology in general. Even if this is not the case, if we analyze the effect sizes of impulse responses in different moments, e.g. 1st quarter, 12th month, etc., and if the conclusions taken from each set do not comply with each other, then it can become difficult to arrive at solid conclusions. 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 not sufficient to form a true effect, and what seemed a possible bias – the generality of the funnel plots looked heavier or skewed to the right – was not corroborated by the PET results (once again focusin only on the published studies sample).

To further screen for possible publication bias in the effect sizes, we tried another scenario in which we added a dummy variable to the models that comprehend all the effect sizes (published and unpublished). The dummy assumes "1" if the effect size is reported in a study by an author associated with the Bank of Japan and assumes "0" otherwise[52]. In this case, the purpose was to assign a specific weight that represents publication bias directly related with the given study characteristic embodied by the dummy variable. According to the methodology applied by Doucouliagos and Stanley (2009), the variables added to the standard model[53], intended to screen for publication bias, now are jointly interpreted with the intercept and the overall bias is the net value of their sum. Similarly, and according to the same authors, if we were to deconstruct the true effect given out by the standard model, we could add to the $1/Se$ (or its proxy) other variables that stand for methodology, data or other study specification types[54], which could help to understand whether these specification differences affect the results

[52] For the sake of simplicity, at the tables, we coin the term "Model with Dummy" to refer to the described specification.

[53] For standard model we mean the MLME or OLS versions where $\sqrt{n_{ij}}$ is the only explanatory variable.

[54] Still on the same authors, variables that deconstruct the (eventual) publication bias are simply put into the model, whereas added variables that intent to deconstruct the true effect are divided (weighted) by their associated standard error (those reported in the studies).

reported in the literature or not. Here, we opted to make just a simple approach for the aforementioned scenario due to the lack of an adequate number of observations.

From all the results obtained[55] there is no evidence of normally distributed errors at 1%, therefore no conclusion can be withdrawn.

Finally, as a closing remark for this section, we do not proceed with any correction of the true effect based on the findings on the publication bias since there is no evidence of these effects. This common procedure in meta-analysis would consist in regressing a t-statistic subtracted by the publication bias (should be the case that this one was statistically significant in the previous exercise) against $1/Se$ (or its proxy)[56].

Table 5-1: Main PET Results

	Maximum Magnitude When Significant		1st Quarter		12th Month		24th Month	
	MLME – All studies	Cluster OLS – Published Studies	OLS – All Studies **(1)**	MLME – Published Studies	MLME – All studies	MLME – Published Studies	MLME – All studies	MLME – Published Studies
β_1 (Intercept)	-21.608 (69.127)	17.822 (55.738)	-16.759 (29.741)	152.878 (106.237)	32.589 (45.290)	76.685 (54.313)	11.378 (54.842)	-64.118 (59.316)
$\beta_0(\sqrt{n_{ij}})$	8.325 (5.446)	2.463 (4.306)	3.215 (3.221)	-16.943 * (10.137)	-0.199 (4.333)	-7.271 (5.205)	2.553 (5.164)	7.646 (5.681)
LR test (p-value)	0.0000 ***	-	-	0.0000 ***	0.0042 ***	0.0000 ***	0.0000 ***	0.0003 ***
ICC ratio	0.763	-	-	0.977	0.363	0.865	0.490	0.879
Jarque-Bera test (p-value)	0.000	0.162	0.000	0.019	0.000	0.195	0.000	0.026
Obs.	48	26	55	25	46	25	46	25

[55] The models with dummy were conducted for all the sets of effect sizes – maximum magnitude, 1st quarter, 12th month, etc.; the outputs are available in tables 1 to 4, Annex II.

[56] As it was already corrected for bias, this model does not include an intercept (Stanley, 2005).

Nr. Clusters	18	9	-	7	13	7	13	7

Notes: *,**,*** if statistically significant at 10%, 5% and 1% confidence level, respectively. We reject the null hypothesis of normally distributed errors whenever Jarque-Bera's test p-value is less than 1%. Parameter's standard errors in parenthesis. The null hypothesis of the LR test states that MLME model is equal to a simple linear model. ICC ratio ranges from [0;1], 0 stands for no correlation within clusters (no variance to explain at cluster level) and 1 stands for maximum correlation (no variance at the individual level, i.e. the observations within clusters have the same value).

(1) OLS instead of the MLME version, since in this latter version the ICC is 0.

6. Meta-probit Estimation

As explained in the previous section, one way to conduct a meta-analysis is 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, thus trying to understand whether there is a significant relationship between the magnitude of the average effect size and each one of those elements. Because we find some degree of uncertainty in regards to the precision of the collected effect sizes which serves as dependent variable – the IRF's magnitude levels – we chose to perform a type of meta-analysis based on a probit regression. This type of model uses the same set of explanatory variables included in the extended PET (linear) model to answer a different kind of question: does a given study characteristic, e.g., the type of output variable used in the studies' framework, affects the probability of a study reporting an output estimation (the effect size), based on a monetary policy shock applied in an impulse response function, in terms of its (overall) sign and significance? As we can see, the former question uses no longer a dependent variable based on a real-valued variable but a categorical one, with the information that we gathered and termed as overall shock's signal. Ideally one would like to use an accurately reported quantified variable (effect size) to perform a meta-analysis – but the information gathered in studies, concerning the IRF's sign and significance, was more readily available in the papers, than the IRF's magnitude[57]. One important detail can be reported back to the descriptive statistics: because there are so few overall IRF's negative effects registered – 2.88% of the total[58]

[57] Recall from sub-chapter 3.2.4, regarding the magnitude of the output estimation: "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." Still in the same sub-chapter, see Table 3-5 to recall how the overall effect of the monetary policy shock in the output was coded.

[58] Being the rest of the stats: positive effects – 50% and non-significant – 47.12%, in 104 observations (Table 9-2, Annex I).

– we perform not an ordered probit, that 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 under the same category. Thus the dependent variable can be read as positive and significant if "1" ("success"), or "0" otherwise (as in Kluve, 2016). The preference for a probit over a logit doesn't matter much but since we are using Stata as the main source for estimations, this software provides a command to test for heteroscedasticity in probit models which is an important element to claim for unbiased coefficient estimates.

6.1 Further Variable Treatment

First and foremost, because we build our own database, we are able to create variables that are intended to not only be informative but also try *a priori* to avoid a number of possible numerical problems that affect negatively the estimation process. The first concern, when constructing categorical variables, is not to use (sub-)categories that are not representative or are very close to a zero cell count – all "1" or all "0". Also, 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, accordingly. There is no explicit justification when to consider a category representation small, but variables with a sub-category with 10% or fewer observations were subject to re-categorization[59]. Table 6-1 presents the resume of the process as well as some other explanatory variables that have been changed or created from a previous set of variables, and available information in our database. The first variable in this table, the Journal Impact variable, was based on the idea presented in Koetse *et al.* (2009) and ranks the effect sizes into three categories – A, B, and C – that were built resorting to the SCIMAGO econometric journal impact ranking in the following manner: 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 also if it is a published study which hasn't been considered in SCIMAGO. Finally, the effect size ranks "C" if it

[59] To the exception of SCIMAGO based ranking variable (sub-category rank A with 8.65%) and the monetary base variable that includes the money supply category (7.69%), that were used in the first stage of probit analysis. See Table 6-1.

was taken from non-published studies. With this ranking variable, we try to assess whether journals with greater impact tend, or not, to report positive significant estimates (a form of publication bias screening). The *Number of Observations of the Analysis' Timeframe* is used as a quantitative variable instead of having two dummies and a reference group[60]. Another version of this latter variable was created where the number of observations falls under ordered intervals of 20 observations[61]. Each of these two last referred variables is used one at a time in the probit model. This variable in intervals was created because using the number of observations per *se* may not be suitable for a posterior marginal analysis of the probit model. We expect, *a priori*, that an increase of one observation will not cause a steep change in the predicted probability of success, at the same time, the variable in intervals also loses some degree of accuracy if the average number of observations within each interval is different from each other. The *Variable that Measures the Output* also had too many sub-categories which were reduced in two different ways, to be tried alternately: the first specification distinguishes variables that are in levels from those that are in differences (reference group); the second specification distinguishes industrial output variables from all other forms. The *Monetary Variable* will be used in the model as originally intended, distinguishing money base from money supply variables and NA cases, but will also alternate in the probit model with the specification that distinguishes solely the money base against "others" (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, with the drawback of removing interest rates shocks out of the inferential conclusions. Confidence Intervals categories that do not correspond to 1 SD width were aggregated into one category. Furthermore, the Interest Rate of Reference and Price Level are simply included in the model in a binary fashion, "present or not present", instead of discriminating every sub-category with dummies. Finally, effect sizes based on daily observations are jointed with monthly based effect sizes.

[60] Recall that the original variable has three categories: lower than 50 Obs, between 50 and 100 Obs, and higher than 100 obs.
[61] The intervals of observations follow this logic until the highest number of observations is met:]0;20],]20;40],]40;60], …,]320;340]. Each interval is then coded from the smallest, 1, to the last, 17.

Table 6-1: New Variables and Re-categorized Variables for Probit Estimation

Variable	Description	Original Categories – frequency in %	Modification	Final Categories – frequency in %
Journal Impact Variable	SCIMAGO Econometrics classification based ranking: A – Q1 AND Q2; B – Q3 and Q4 and non-rated, C – non-published	-	-	A – 8.65 B – 36.54 C – 54.81
Periodicity of the Time Series	-	Daily – 1.92 Monthly – 53.85 Quarterly – 44.23	Daily and Monthly data in the same category	Daily and Monthly – 55.77 Quarterly – 44.2
Nr. of Observations (Obs)	Number of Observations used to produce the effect size	L50 – 12.50 B50100 – 43.27 H100 – 44.23	Quantitative variable instead dummy scheme	-
Nr. Observations in intervals (Int. Obs)	The previous observations are fit into ordered intervals of 20 variables. From interval 1 –]0;20] Obs; until interval 17 –]320;340] obs.	-	-	-

Variable(s) that measures the output 1	GDP if - GDP; or Real GDP. GDP growth if - GDP growth; Real GDP growth. OG – Output gap (or proxy) if - GDP gap; Investment gap; Output gap. IO – Industrial Output if - Total Industry Activity Index; Industrial Production Index; Industrial Production; All Industry Activity Indices; Real Production Index; Electric Power Consumption; Economic Activity Factor. UR – Unemployment Rate.	GDP growth and OG – 25.00 GDP – 12.50 IO – 61.54 UR – 0.96	1 if the output variable is in levels (IO and GPD); 0 otherwise (variable in differences – GDP growth, OG and UR)	Output in levels – 74.04 Output in differences – 25.96
Variable(s) that measures the output 2	Alternative to the former form: 1 if Industrial Output; 0 otherwise.	IO – 61.54; Otherwise – 38.46	-	-
Monetary Variable	MB – Money Base; MS – Money Supply	NA – 10.58 MB – 81.73 MS – 7.69	1 if MB; 0 otherwise (NA + MS)	MB – 81.73 Other – 18.27
Interest Rate of Reference (IRR)	3MIR – 3-month interest rate CR – Call rate RR – Repo Rate	NA – 46.15 3MIR – 2.88 CR – 49.04 RR – 1.92	1 if IRR; 0 otherwise	IRR – 53.85 NA – 46.15
Price level (or Proxy) (PL)	CPI – CPI IR – Inflation Rate CCIG – Core CPI inflation Gap GDPD – GDP deflator	NA – 14.42 CCIG – 1.92 CPI – 57.69 GDPD – 0.96 IR – 25.00	1 if Price Level; 0 otherwise	PL – 85.58 NA – 14.42

Type of Shock (1)	OTS – Other Type of Shock SMB – Shock to the Monetary Base SMS – Shock to the Money Stock SSTIRR – Shock to a Short-term Interest Rate of Reference	OTS – 6.73 SMB – 52.88 SMS – 27.88 SSTIRR – 12.50	OTS aggregates with SSTIRR	SMB – 52.88 SMS – 27.88 OTS and SSTIRR – 19.23
Confidence Intervals	68PCI if 68% Confidence Intervals; (-/+) One -standard Deviation Confidence Bands; 16th and 84th Percentile Confidence Bands 90PCI-1090PCB if 90% Confidence Intervals 10th and 90th Percentile Confidence Bands 95PCI if 95% Confidence Intervals; (-/+) Two-standard Deviation Confidence Bands	NA – 11.54 68PCI – 41.35 90PCI-1090PCB – 18.27 95PCI – 28.85	90PCI-1090PCB aggregates with 95PCI in a single category	NA – 11.54 68PCI – 41.35 “Other Widths” – 47.55

6.2 The Probit Model and Preliminary Procedures

The formalization of the probit model is the following:

$$Pr(y_i = 1|x_i) = \Phi(x_i b) \qquad (4)$$

$i = 0, 1, 2, \dots,$ 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 and/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, pp. 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 to build the model on a stepwise basis that may allow for more statistically robust model estimations. Another numerical problem caused by the inclusion of too many variables in the model – the complete or quasi-complete separation –, happens whenever a combination of explanatory variables are able to determine perfectly (or almost perfectly) the pattern of the outcome binary variable[62]. This modelling strategy starts with a preliminary analysis of the variables to be included in the model. One variable at a time, or a categorical group at a time[63], we include these in a probit, modeling the binary outcome variable described in (1), and we examine whether the variables are significant or tend to be zero, 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, where the data was clustered by paper of origin[64]. Because the coefficients' values of explanatory variables can change when interacting with other variables, we only exclude variables at this stage, when both standard and cluster estimations report *p*-values over 0.2. This looser criterion also permits to include in the next stage variables that contain explanatory information that could be prematurely excluded otherwise.

Table 6-2, which reports the *p*-values of the coefficient's z-statistic of the univariate models, already contains some important information for the whole scope of the present

[62] An example of a situation of complete separation, based solely on a model with two explanatory variables of our database, would be if an effect size was always statistically significant and positive, given that the output variable was an Industrial index and at the same time, the observations that originated effect sizes were based on quarterly data.

[63] Hosmer and Lemeshow (2000, pp. 38) recommend that whenever a categorical group is composed by more than one dummy, the whole dummy scheme should be modeled and tested altogether during the UNIVARA process. Should it be the case that at least one of the dummies is not significant, the same authors propose that the whole category must be re-configured in order to provide significant dummies; the point being, increasing the number of degrees of freedom when resorting to the Wald statistics.

[64] Throughout the meta-probit – Section 6 – 104 observations and 30 clusters were used to fit the various models.

study. To begin with, there is a group of variables that was excluded (“Ex”) at the next stage of the analysis, from which we can infer that the probability of an effect size being significant and positive does not depends on:

- Whether the effect size is taken from a published or unpublished study.
- The studies’ impact ranking[65].
- Whether one of the author’s 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 in which the effect size is based on, is in levels or in differences.
- The inclusion of an interest rate variable or an exchange rate variable in the model of a study’s framework.

Although there is some ambiguity regarding the statistical significance of the type of monetary variable (MV2 with the lowest p-value at 14.6%), this dummy was not overruled from the analysis at this point. In alternative to the original two-dummy scheme, the monetary variable was tested in a single dummy form – “1” if money supply, “0” otherwise – in order to understand whether this specification proved its significance. The result turned out to be ambiguous since there is statistical insignificance for the cluster version, which will lead to the alternative testing of these two variable schemes in the next stage. In a similar manner, 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 money supply. The results from the individual tests at this latter dummy scheme did not perform as well: the dummy variable that discriminates shocks applied to the monetary base (Shock a) almost violates the 0.2 threshold criterion. Nevertheless, besides the dummy Shock c, which isn’t quite informative on its own, both dummies Shock a and Shock b were tested individually in the next stage of the analysis, as an alternative to the original scheme. The individual tests related to the use of different types of confidence intervals were found highly insignificance when discriminating only 1 SD confidence intervals

[65] Although in the standard version the dummy variable that signaled A ranked studies was near 5% significant, the addition of the whole dummy scheme of the Journal Impact Ranking in the next stage of the analysis did not produce any relevant conclusions.

(CI1), which may suggest that the probability of success does not depend on this standard band width. Despite this, the dummy CI2 is statistically significant, at least in its non-cluster version, and therefore was not dropped. From a correlation matrix[66] of the variables presented in the Table 6-2, it was also possible to identify some cases of high collinearity, namely between QE and Year, and between IO and Periodicity. This will require the fitting of these variables in separate specifications, in order to less precision, and thus the exclusion of variables based on the lack significance of their coefficients.

Table 6-2: P-values of the Coefficient's Z-statistic taken from Univariate Probit Models

Next Stage Analysis	Variable's Code Name	Variable's Description	*P*-values of the coefficient's Z-statistic	
			Standard Probit	Clustered by Study
Ex	Publication One	1 if published; 0 otherwise.	0.168	0.455
Ex	Bank of Japan	1 if authors are associated with the BoJ; 0 otherwise.	0.513	0.735
Ex	Interest Rate	1 if the model has an IR; 0 otherwise.	1.000	1.000
Ex	Exchange Rate	1 if the model has an Exchange Rate; 0 otherwise.	0.658	0.696
Ex	Empirical Method	1 if model is VAR-VEC; 0 otherwise (TVP-BVAR-Switching).	0.156	0.476
In	QE	1 if studies' analysis timeframe comprehends only the 1st QE period; 0 otherwise.	0.000	0.019
In	Periodicity	1 if quarterly data; 0 otherwise.	0.006	0.146

[66] See the correlation matrix at Table 11-1, Annex III.

In	Year	Year of the study's publication.	0.000	0.038
In	Obs	Number of observations used to produce the effect size.	0.051	0.164
In	Int. Obs	Nr. of observations ordered by intervals.	0.062	0.193
Ex	Outputvar	Variable that measures output: 1 if variable is in levels; 0 otherwise (differences).	0.264	0.504
In	IO	Variable that measures output: 1 if the variable is an Industrial Output; 0 otherwise.	0.005	0.099
In	Price	1 if the model has a price level variable; 0 otherwise.	0.003	0.029
In	Bond	1 if the model has a bond yield variable; 0 otherwise.	0.006	0.029
Ex	(i.abc 0)	Group of reference is the C group classification – non-published studies.	-	-
	i.abc 1	Ranked 1 if the published study has a classification of B.	0.402	0.662
	i.abc 2	Ranked 2 if the published study has a classification of A.	0.063	0.204
In	(MV 0)	Group of reference is "NA".	-	-
	MV1	1 if the monetary variable in the model is considered money supply; 0 otherwise (other or no monetary variable).	0.028	0.030

	MV2	1 if the monetary variable in the model is considered money base; 0 otherwise (other or no monetary variable).	0.146	0.237
In	MV1	Testing MV1 individually.	0.077	0.191
In	(CI0)	Group of reference is "NA".	-	-
	CI1	1 if the confidence interval has a 1 SD width; 0 otherwise.	0.046	0.021
	CI2	1 if the confidence interval other width besides 1 SD; 0 otherwise.	0.010	0.000
Ex	CI1(individually)	Testing CI1 individually.	0.842	0.920
In	CI2 (individually)	Testing CI2 individually.	0.077	0.311
In	Shock a (SMB)	1 if the shock is applied to the money base; 0 otherwise.	0.005	0.071
	Shock b (SMS)	1if the shocks applied on the money supply variable are the reference group.	-	-
	Shock c (Others and SSTIRR)	1 if the shock is applied to other variables besides SMB and SMS; 0 otherwise.	0.002	0.012
In	Shock a	Testing shock a individually.	0.169	0.401

In	Shock b	Testing shock b individually.	0.001	0.033
Ex	Shock c	Testing shock c individually.	0.048	0.102

1) The first column – Next Stage Analysis – states whether the variable is included (In), or excluded (Ex) of the next stage of the analysis.
2) 104 observations were used in all estimations. 30 clusters were used to produce robust standard error models.

6.3 Stepwise Variable Selection Analysis of the Probit Model

The second stage of the analysis uses a stepwise backward variable selection, as part of the model's modelling strategy. Its purpose is related with the fact that after the UNIVARA stage there is still a considerable number of variables to add into the model and there is no theory that supports a hierarchy of relevance in regards to what variables should be chosen over others, given that too many variables weakens the model's statistical properties. This type of procedure works as an algorithm, and for this purpose we used the designated Stata command. Using a stepwise backward selection entails that the model is built by fitting initially all the variables – the full model –, then Stata starts a search for the removal of the least significant coefficient, given a pre-determined threshold. If a variable is removed, the algorithm re-estimates the model without the excluded variable and repeats the same search for removal. The process is repeated in a backwards fashion until all variables are under the significance threshold. For the present case, the significance level used for the coefficient's Z-statistic was set at 8%, a value that allows the inclusion of border line variables that exceed the conventional 5% significance level. The list of variables that is designated to be tested by the Stata algorithm is the one previously cleared at the UNIVARA stage. Like in the UNIVARA analysis, the stepwise estimation process was performed using standard errors and cluster robust standard errors and also had into account the collinear cases early reported by testing those variables separately. It is important to mention that the number of estimated coefficients for the standard model and its cluster robust counterpart may not match. This happens because although the list of variables used at the Stata's stepwise algorithm is the same, the procedure may produce distinct outcomes, i.e., different variables may be excluded during the process. For sake of comparison, Table 6-3 reports both standard and cluster robust SEs for the same specification.

The most relevant results produced from the stepwise estimations are shown in Table 6-3. These results were selected from a battery of estimations performed using alternative specifications as explained earlier. These results were chosen having in consideration the performance of the Pearson's Chi-squared goodness-of-fit (GoF) test, which infers on the ability to fit the probit model well to the data[67]. Furthermore, because the stepwise algorithm doesn't take into account how good a model is in terms of goodness-of-fit, we did a subsequent round of estimations that consisted in fitting stepwise models presented in Table 6-3 but removing or replacing a variable at a time. The objectives were to: a) improve the goodness-of-fit of a previous model; b) reach alternative model specifications, in terms of selected variables that were not produced via stepwise. From this round, the models that were considered are: model 1-b, in which the variable Obs (and Int. Obs) were removed from model 1-a, increasing substantially the goodness-of-fit; and model 4-b, similar to model 4-a, with the substitution of IO by Periodicity, maintaining the statistical significance of all parameters and a goodness-of-fit well over a 10% level. Models 1-b, 2 and 5 may seem redundant since they report similar results to those in model 1-a, but they have been included in the analysis nonetheless, because when testing for heteroscedasticity bias (see chapter 6.5), it may be useful to test alternative specifications, since the likelihood convergence of those models is difficult to attain.

From a general perspective, the estimation results presented in Table 6-3 report standard and clustered SEs that are not overwhelmingly high. According to Hosmer and Lemeshow (2000, pp. 135-141), suspiciously high coefficient values or/and standard errors may be caused by the numerical problems earlier enunciated: quasi- or complete separation and collinearity between variables. Moreover, the robust SEs maintain the significance of the parameters, when compared with standard models. In terms of selected variables, adding both CI1 and CI2 in the model did not produced any relevant estimation, neither when CI2 was considered alone[68], thus we can infer that the use of different intervals of confidence band widths, do not affect the probability of attaining a positive and statistically significant effect size. The same can be said when the effect size is controlled for the year of publication of the paper of origin, and whether the

[67] The Pearson GoF test is more reliable when the number of covariate patterns is lower than the number of observations Hosmer and Lemeshow (2000, pp. 144). For the models presented in Table 6-3, the number of covariate patterns is always considerably lower than the number of observations.

[68] Remind that CI1 (effect sizes with 1 SD band width) is highly statistically insignificant when individually tested at the UNIVARA stage, thus it has been ruled out of the stepwise analysis.

shock applied on the money supply (shock b) or on other types (shock c). We can also state that because we chose to use a stepwise estimation procedure that removes insignificant coefficients at 8% level, the variables that made it through into Table 6-3 are those that affect the estimations (effect sizes) reported by the studies:

QE – there is evidence in six of the seven models presented in Table 6-3 that studies focused in the first Japanese quantitative period (2001-2006) have a lower probability to find evidence of a positive and statistically significant output effect than other QE periods.

Periodicity – model 3, which is similar to model 2 but replaces IO by Periodicity due to collinearity, suggests that studies that use (daily or) monthly data over quarterly data have an increased probability of producing positive and significant effect sizes. Model 4-b gives the same impression.

IO – all models presented in Table 6-3 that includes the variable IO suggests that if a study uses an industrial output as its explanatory variable (output proxy) the probability of estimating a positive and significant result is higher compared to the use of other output variables.

Price Level – the price level is included in model 4-a (and 4-b) presented at Table 6-3 which resembles model 2. The use of a price level variable as an explanatory variable is more likely to produce a positive and significant effect size than not using it.

Bond – in all models, the common determinant is the bond (yield) variable in which whenever a study includes a variable of this kind the predicted probability of reaching a significant and positive size-effect is lower than not including it.

MV1 and *MV2* – to the exception of models 4-a and 4-b, both variables that distinguish from different types of monetary variables 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 originates a lower probability of producing a significant and positive effect size than using other kind of non-monetary variable as the Japanese policy tool, e.g., interest rate of reference.

Shock a – from the several estimations, only the dummy variable that distinguishes shocks to the monetary base (shock a) from any other type of shocks has been included in three of the reported models – 1-a, 1-b, and 5. There is evidence that the use of a monetary base tool rather than any other type of monetary variable increases the probability of producing a positive and significant effect size.

Obs and *Int. Obs* – these two variables which have been used in alternation, show similar estimations across the panel of probit models here analyzed; taking evidence from models 1-a and 5, more observations increase the probability of reaching a positive and significant effect size.

Table 6-3: Meta-probit Estimation Results

	Model 1-a	Model 1-b	Model 2	Model 3	Model 4-a	Model 4-b	Model 5
QE	-0.753 (0.344)*** (0.317)***	-0.598 (0.319)** (0.332)*	-0.715 (0.336)*** (0.306)***	-0.803 (0.302)*** (0.374)***	-0.612 (0.299)*** (0.308)***	-0.648 (0.299)** (0.318)**	-
Periodicity	-	-	-	0.851 (0.297)*** (0.437)***	-	0.702 (0.287)** (0.353)**	-
Industrial Output (IO)	0.709 (0.355)*** (0.299)***	0.900 (0.329)*** (0.339)***	0.975 (0.321)*** (0.374)***	-	0.667 (0.296)*** (0.309)***	-	0.845 (0.328)*** (0.387)***
Price level	-	-	-	-	1.315 (0.496)*** (0.567)***	1.25 (0.481)*** (0.613)**	-

Bond Yield	-1.500 (0.494)*** (0.422)***	-1.510 (0.478)*** (0.485)***	-1.440 (0.481)*** (0.544)***	-1.425 (0.477)*** (0.518)***	-1.115 (0.390)*** (0.321)***	-1.170 (0.395)*** (0.305)***	-1.745 (0.480)*** (0.426)***
MV1	-2.607 (0.711)*** (0.532)***	-2.225 (0.617)*** (0.470)***	-1.697 (0.557)*** (0.369)***	-1.530 (0.575)*** (0.379)***	-	-	-2.736 (0.677)*** (0.564)***
MV2	-2.375 (0.783)*** (0.787)***	-1.725 (0.686)** (0.797)**	-1.762 (0.682)*** (0.809)***	-1.613 (0.694)** (0.793)**	-	-	-2.098 (0.749)*** (0.764)***
Shock a	1.020 (0.384)*** (0.486)**	0.756 (0.345)** (0.433)*	-	-	-	-	1.080 (0.364)*** (0.532)**
Obs(2)	0.005 (0.002)*** (0.002)**	-	-	-	-	-	0.005 (0.002)** (0.002)**
Int. Obs	0.113 (0.044)** (0.053)**	-	-	-	-	-	0.097 (0.042)** (0.050)***

Constant	1.179 (0.650)* (0.354)***	1.425 (0.565)** (0.345)***	1.39 (0.555)*** (0.371)***	1.428 (0.587)** (0.433)***	-1.166 (0.531) (0.585)**	-1.060 (0.495)** (0.649)*	0.997 (0.619) (0.379)***
Log-likelihood or – pseudolikelihood	-47.105119	-50.920927	-53.361444	-53.97263	-55.210395	-54.751231	-49.688369
GoF's *p*-value(1)	0.0594	0.3192	0.2873	0.2040	0.5453	0.1576	0.2622
LR test Homo vs Hetero	-	-	-	-	0.3225	0.8021	0.0064
Note 1	Int. Obs GoF's *p*-value: 0.0404	"Manually" Removing Obs from Model 1.	-	-	-	Model fitted "manually" from model 4-a – Periodicity instead of IO.	The algorithm excluded the year variable.

Note 2	Obs with similar estimation results to Int. Obs. Obs's coef. = 0.005.	-	-	-	-	-	*Obs* with similar estimation results to Int. Obs. Obs's coef. = 0.005.

(1) GoF's p-value is taken from a Pearson's Chi-squared goodness-of-fit test. One does not reject that the data fits the probit well if the *p*-value is over 5%.
(2) Obs coefficient's estimation and standard error are part of an alternative specification, which is not reported here, where the only variable change is Obs by Int. obs.
(3) A total of 104 observations were used.
(4) First brackets report standard errors for the standard model; second brackets report cluster robust standard errors at the study level (30 clusters were used).
(5) ***,**,* when the coefficient's estimate is and statistically significant at 1%, 5%, between 5 and 8%; being the only exception model 5 (standard version) constant term (*p*-value = 10.2%).
(6) LR test Homo vs. Hetero, is the Stata's test – "Likelihood-ratio Test of $\text{lnsigma}^2 = 0$" – that checks the overall significance of the regressors within the probit's variance, that are associated to the index variables. A test's *p*-value under 5% suggests that the heteroscedastic probit model is less biased, in regards to the parameters, than the homoscedastic standard probit model.

6.4 Adjusted Predictions and Marginal Effects

To better understand the previous results, we obtain adjusted predictions and marginal effects extracted from the models provided in Table 6-3. This analysis excludes model 5 since the findings reported in chapter 6-5 suggest that it may be heteroscedastic. First, we explore how the only continuous variable that has been included in the model – Obs (or in alternative Int. Obs) – predicts the outcome variable, i.e. studies that report an effect size that is positive and statistical significant. The model specification used is 1-a (Table 6-3). For this adjusted prediction we set Obs at three different values – 50, 100, and 150 (for Int. Obs we use the correspondent intervals 3, 5, and 8[69]) and all other variables at their sample mean. From the results depicted in Table 6-4 both variables return the same result[70].

[69] Intervals of observations are as follows: 3 is]40;60], 5 is]80;100], and 8 is]140;160].
[70] This analysis would benefit from controlling simultaneously the number of observations – Obs or Int. Obs – and the periodicity of the data – monthly or quarterly – but no model included them together in the stepwise stage of the analysis.

Table 6-4: Adjusted Prediction of the Outcome Variable for Different values of Obs and Int. Obs

Number of Observations	Prob($y = 1$)	Intervals of Observations	Prob($y = 1$)
50	0.332784	3	0.348016
100	0.445598	5	0.434656
150	0.563031	8	0.569339

1) Adjusted prediction derived from the model 1a. All other variables set at their means.

One oddity of the adjusted prediction of outcome presented in Table 6-4 relates to setting all other variables, except Obs, at their sample means. This results in an awkward interpretation because those variables are dummies. To set an example, if we fit a continuous variable, such as Year (of Publication), we have an average year of publication. If it is a dichotomous variable at their mean, the interpretation is something like: Industrial Output (IO) set at 61.5% – its average value, the interpretation is that 61.5% of the observations reported the use of an Industrial Output variable. Having said this, the interpretation using the specification "at means" cannot presume that we are "comparing the average study" and just changing the values of Obs. To dodge this interpretation caveat, the next marginal effects analysis gives the change in the predicted probability of success ($y=1$) associated with the change from "0" to "1" at one dummy[71] and setting all other variables at a given value; in this case we used the actual observed values of the variables, an approach designated as Average Marginal Effects. The best way to explain this approach is to describe the process of calculating a single average marginal effect. Suppose we want to know the change in the predicted probability of success when the price level is considered (used in the model that produces the output estimation) in opposition to not being used, that is, changing the dummy from "0" to "1" in the probit specification. To do this, a series of steps can be taken in order to give an average value of the marginal effect:

1) Start by calculating the predicted probability of the "first" effect size, that is, by setting the price level as not in the model, "0", and using the actual observed values of

[71] To the exception of the continuous variables Obs and Int. Obs.

the other variables[72] in x_i with $i = 1$. Repeat in the same manner for the other 103 cases, since we have 104 cases/observations. Then take the average.

2) Proceed as in 1) but now assuming that the price level is used, “1”, in all 104 cases and then take the average of those predicted probabilities.

3) The overall average marginal effect of using a price level variable is given by subtracting the average predicted probability of not using the price level from the average predicted probability of using it.

The reasoning behind this process is to not think about computing marginal effects using two sub-populations – those that use price level and those that don’t. As in both cases, all other variables are the same at their observed values, then the only element that changes is the exclusion or not of the price level. Those two scenarios produce different predicted probabilities whose difference is our (averaged) marginal effect.
Table 6-5 presents the average marginal effects (AME). In order to ease the interpretation of the marginal effects, whenever the variable in question is dichotomous the interpretation can be reversed. One way to analyze these effects is to compare their values within the same model and then across models: the AMEs with the highest 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 a study will report a positive and significant effect size decreases immensely, roughly around 45 and 65 percentage points[73]. 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 behavior of the impulse response functions when a monetary policy shock exists, and thus affecting the output estimation in terms of its sign and significance. A noticeable discrepancy is given by the different signs in MV2 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)[74]

[72] Recall the formalization of the probit model, (4), given at the beginning of the sub-chapter 6.2.
[73] To ease the interpretation of the results in this sub-analysis, we will abstain to imply the standard-errors variation, and instead, describe how the AMEs varies from model to model.
[74] Recall that *Shock a* stands for a shock in the money base (in opposition to other monetary variables).

then the probability increases between 26.7 and 20.6 percentage points (according to model 1-a and 1-b, 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 around 33 and 25 negative percentage points. Regarding the other variables, the magnitude of the AMEs of including a price level variable or not including a bond yield are higher (roughly 38 and 35 percentage points more, respectively), and thus more relevant than the choice of an industry index as an output proxy (averaging across models an AME of 24 percentage points). The magnitude of the AMEs for the specification that studies the 1st QE period or chooses quarterly data are lower than those previously discussed, but still quite high. The exception to the very high magnitudes found is the case of the continuous variable In. Obs. Recall that the purpose of this variable was to be used instead of Obs in order to measure a change in one unit that represents an increase of 20 observations. Its average marginal effect is of a 2.8 percentage point increase in the probability of studies reporting a positive and statistically significant effect size.

Table 6-5: Average Marginal Effects

	Model 1-a	Model 1-b	Model 2	Model 3	Model 4-a	Model 4-b
QE	-0.190 (0.081)**	-0.163 (0.083)**	-0.207 (0.081)**	-0.235 (0.079)***	-0.182 (0.084)**	-0.192 (0.082)**
Periodicity	-	-	-	0.249 (0.075)***	-	0.208 (0.078)***
Industrial Output (IO)	0.179 (0.084)**	0.245 (0.080)***	0.282 (0.079)***	-	0.199 (0.081)**	-
Price level	-	-	-	-	0.392 (0.132)***	0.371 (0.129)***
Bond Yield	-0.378 (0.107)***	-0.412 (0.113)***	-0.417 (0.123)***	-0.418 (0.124)***	-0.332 (0.102)***	-0.346 (0.102)***
MV1	-0.658 (0.141)***	-0.607 (0.134)***	-0.491 (0.140)***	-0.449 (0.152)***	-	-
MV2	-0.600 (0.171)***	-0.471 (0.170)***	-0.510 (0.178)***	-0.473 (0.188)**	-	-
Shock a	0.267 (0.085)***	0.206 (0.086)**	-	-	-	-
Int. Obs	0.028 (0.010)***	-	-	-	-	-

1) Delta-method standard-errors in brackets. *,**,*** if statistically significant at 10%, 5% and 1% confidence level, respectively.

6.5 Heteroscedasticity Analysis

One problem in probit and logit models that may cause inconsistency of the MLE is heteroscedasticity (Davidson & MacKinnon, 1984). In the former probit model (4), at the beginning of sub-chapter 6.2, is assumed homoscedasticity, which is formalized by imposing a normal CDF with constant variance of 1. The problem of this assumption is that it is often too optimistic and not verified. Heteroscedasticity can be defined as a non-constant variance that depends on the variable(s) in the index function. In regards to the present study, one could ask whether the probit estimations reported earlier at Table 6-3 are biased due to heteroscedasticity, or not; and also, what one could conclude differently from a heteroscedastic probit model. Stata provides a command that permits to fit a maximum-likelihood heteroscedastic probit model which can help answering the previous questions[75]. The following formalization of the model is based on the Stata Base Reference Manual[76]:

Recall the homoscedastic scenario as in (4):

$PR(y_i = 1|x_i) = \Phi\left\{\frac{x_i b}{1}\right\}$, where Φ is a Gaussian CDF with constant variance, $\sigma_i^2 = 1$, for all i.

With heteroscedasticity, the variance varies through a multiplicative function of the index variables:

$$\sigma_i^2 = \{exp(z_i\gamma)\}^2$$
$$z_i = (z_{1i}, z_{2i}, \ldots, z_{mi}).$$

Where z_i, in the words of Harvey (1976), "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". The heteroscedastic probit model yields a probability function of success of the following type:

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

For this analysis, we fit[77] heteroscedastic probit models with the same variables as in the previous models reported in Table 6-3 using Stata's commands. The estimation of this type of model is known to have difficulties in convergence and indeed it only succeeded

[75] The RESET test could also be employed to detect functional form problems such as heteroscedasticity and sample related misspecifications.

[76] To check the formalization into more detail, see Stata Base Reference Manual – Release 13 – "Heteroskedastic probit model".

[77] Regular standard errors were specified; the use of cluster robust standard errors was not because Stata Base Manual refers to this option as inefficient for this type of model.

for models 4-a, 4-b and 5. For the remaining models (1-a, 1-b, 2, and 3) no convergence was achieved after 16000 iterations[78]. Resorting again to Table 6-3, the LR test checks whether the heteroscedastic version is statistically significant over the previous homoscedastic version, and thus, potentially less biased in terms of the parameter estimates. The complete results for the heteroscedastic probit estimations are shown in Table 11-2, in Annex III. Model 5 presents evidence of heteroscedasticity (LR test with a *p*-value under 1%) but at the same time reports very disproportionate and large coefficients and standard errors, thus suggesting that this model specification is not reliable and might contain numerical problems. Inversely, models 4-a and 4-b present an insignificant LR test, which indicates that there is no reason to suspect of heteroscedastic bias from the previous standard versions. Moreover, none of the model's variance coefficients γ is significant. In resume, the evidence reported earlier at Table 6-3 for models 4-a and 4-b is reliable since the present analysis suggests that they are heteroscedasticity bias free.

6.6 Interaction Effects[79]

The search for interaction effects among variables has the purpose of trying to assess whether the pattern of a dependent variable is somewhat related with, or can be described by, how two (or more) of its explanatory variables relate with each other. Another way to see an interaction effect is to assess if the relation between the dependent variable and one of its explanatory variables depends on the magnitude of another independent variable (Norton *et al.*, 2004). According to the same authors, interaction effects in non-linear models with two 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:

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

[78] It can be specified which index variables will model the variance regressors – γ. This takes place if it is known, or suspected, which variables can cause heteroscedasticity in the first place. Because we do not possess any relevant basis to justify the exclusion of any of the index variables as a possible source of heteroscedasticity, we always include them all in $\boldsymbol{z}_j$.

[79] The analysis of interactions can be preceded by the marginal analysis (sub-chapter 6.4) because it is not possible to calculate marginal effects, at least how you would interpret them for a non-interacted variable, since they are always dependent on the individual component terms which compose the interaction.

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 from.
X, represents all other regressors in the index function.
Φ, has the same properties as in (4).

The interaction for non-linear models is given by the following cross-partial derivative:

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

It is important to make notice of a few features from (6) and (7):

- when the probit is fitted with the interaction term $\beta_{12}\, x_1 x_2$, its statistical significance does not provide the proof of the validity of the effect, neither its sign and magnitude should be regarded, since the interaction effect is given by (7) instead.
- (7) is estimated for different covariate values so that interaction effects may differ in magnitude, sign, and significance. Therefore, the analysis is more complex than in the linear case and relies on the identification of patterns of the relation between the interaction effect and the predicted probability of the probit model.

As the analysis of interaction effects is somewhat burdensome, when analyzing the whole spectrum of interaction effects provided by the calculation of the cross-partial derivatives for all the covariate values, we simplified this analysis to the models that have been proven more reliable before – the heteroscedasticity bias free models 4-a and 4-b (from Table 6-3). First, we searched for the interaction terms that, regardless of their significance, do not inflict statistical insignificance on the other parameters. This was done by fitting an interaction term at a time[80]. Next, we analyzed 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 (performed

[80] We used Stata's command – Inteff – fitting a model with an interaction term a time. The command provided the figures III-a to III-h that regard the calculation of the "interaction effects after probit" and the correspondent "z-statistics of interaction effects after probit". The mean value of the interaction term, plus its standard error and z-stat are also provided by the same command, but were omitted due to the lack of statistical significance of the interaction effects.

by Stata). From the previously described procedures, there were a few cases when the inclusion of the interaction term did not interfered with the statistical significance of the other parameters (see Annex III, figures 11-a to 11-d), but when analyzing the statistical significance in each one of those cases, we find no statistical significance for any of the interaction effects.

6.7 Discussion of some Methodological Choices

One of the implicit aspects of the meta-probit analysis presented in this chapter is that there isn't one single model from which to infer, rather, a number of them that, not contradicting each other plainly, do differ partially in what conclusions one might take. So, if one were to choose the model that was more statistically consistent, and thus, more reliable in terms of what could be inferred, which one would be in the present case? According to the Table 6-6, LR tests suggest that the models with more variables – model 1-a and 1-b – provide a statistically significant better fit (*p*-value under 5% in all cases possible tests) in terms of maximization of the likelihood function; but the same does not happen when the unrestrained models are model 2 and 3, and the restrained ones are model 4a and 4-b. Model 1-a, although performing better in terms of the LR test than the rest, provides the worst Pearson's GoF test (*p*-value: 0.0594), which may indicate that the predicted values produced by the model deviate substantially from the observed values of the outcome variable. Another issue relates with the possibility of heteroscedasticity biased parameters. Since we were only able to fit and analyze some of the former models through this scope, we found out that model 4-a and 4-b were the more reliable. This is crucial in assessing the inferential reliability of a model, because there is no reason not to suspect, that models that did not converged when fitting the maximum-likelihood heteroscedastic probit model, are not biased. Taking into account the significance of the parameters, the highest goodness-of-fit, and the verification of homoscedasticity, Model 4-a emerges as the most reliable to withdraw conclusions. This premise poses a few considerations. On one side, model 4-a is in consonance with other models, regarding what might be inferred from the QE, IO and Bond Yield terms; but on the other side, model 4-a and 4-b, being the shorter models and including Price, do not include MV1, MV2, Shock a or Obs./Int. Obs.; if model 4-a and 4-b can be considered more robust estimation wise, then the evidence of the effect of having a price level variable in the model is stronger than the evidence of the effects of those other variables (MV1, MV2, Shock a or Obs./Int. Obs) that are not found in these models.

Table 6-6: Likelihood Ratio Test Chi-square p-values

	Model 1-a	Model 1-b	Model 2	Model 3	GoF's *p*-value[1]
Model 1-a (7)	-	-	-	-	0.0594
Model 1-b (6)	0.0066	-	-	-	0.3192
Model 2 (5)	0.0022	0.0272	-	-	0.2873
Model 3 (5)	0.0012	0.0135	-	-	0.2040
Model 4-a (4)	0.0012	0.0137	0.0545	0.1156	0.5453
Model 4-b (4)	0.0018	0.0137	0.0955	0.2121	0.1576

1) The number of fitted variables in the model is show in brackets (first column).
2) GoF's p-value is taken from a Pearson's Chi-squared goodness-of-fit test. One does not reject that the data fits the probit well if the p-value is over 5%.

Regarding methodological choices, it is a common concern to deal with the fact that the contribution of studies for the meta-analysis dataset may be uneven, i.e., some studies may report many effect sizes, and others may not, which is the case. To smooth the contributive weight of each study in the estimations of the "meta-model", the literature resorts to several different approaches. One procedure is to use the weight every effect size from a study, with the inverse of the total number of estimates reported by that study or using the standard error associated with each estimate instead. Other is to attribute a weight to the effect size based on the perceived quality/reliability of the publishing journal, or simply choose a fixed number of the "best" reported effect sizes, according to some criteria[81]. In our case, the meta-probit would not benefit from these procedures because it is mainly composed by categorical variables translated in dummies, therefore, the conclusions taken from the meta-probit estimations are subject to the uneven contribution of studies. The only procedure that did account for the correlation of the effect sizes, but did not changed the parameters values based on that fact, was to estimate probit versions with cluster robust standard errors at the study level. As in Section 5, with the PET model, an alternative multilevel mixed-effect probit model could be employed, in order to account for the possibility of within-study correlation – that is affected by the number of effect sizes reported by each study – producing, eventually, different parameter estimations and/or differing in terms of their statistical significance. Once again, a model mainly based on categorical values would not render the use of this alternative model, as beneficial. We opted instead, by the standart probit, since it also provides a series of tools that this latter option does not: the use of the Stata's stepwise algorithm, the calculation of the Pearson GoF test, the comparison between the heteroscedastic corrected probit and the standard probit. Another acceptable criticism of the present analysis is related to the choice of a framework based on a probit model. The problem resides in the fact that the outcome variable recognizes a very strict distinction between effect sizes, which is whether they are positive and significant or not, and thus, disregarding the magnitude of the effect size. In a case like the Japanese, where the effect of monetary policy shocks only results in small, near to zero, positive output growth, it dilutes the importance of distinguishing the positive effect size from the non-significant. In an alternate scheme for a future analysis, one could set the success of the probit's outcome variable indexed to a magnitude threshold given a certain criterion, e.g. $y=1$, if the output growth caused by the shock is significant, positive, and higher or equal to 1% at some point in time, during the shocks length. In this example, the issue is to justify why the 1% increase in output would be a relevant threshold.

7. Conclusion

The present meta-analysis focused on the monetary policy literature that makes use of the vector auto-regressive methodology (in its various forms), to study the impact of the

[81] As referred in sub-chapter 3.1, we dismiss this last solution to avoid selection bias.

quantitative easing programmes set by the BoJ, in the Japanese output growth, and resorts to the analysis of the impulse response functions that describe the output reaction to a monetary policy shock; the main goal of the meta-analysis was to screen for biased reported results in published studies, and understand whether elements that characterize the addressed literature, affect the probability of studies reporting an overall positive and statistically significant effect on output (the effect size). The database built for the meta-analysis is a panel composed by a total of 104 observations collected from 30 studies, containing elements that describe the effect sizes themselves and several other elements – methodological and data related – directly implied on how these effect sizes have been conceived. Furthermore, the database is composed by 47 observations collected from 15 published studies, while the other 57 observations come from 15 non-published studies. From this sample, we registered that 50% of the effect sizes reported an overall positive and significant effect of the monetary policy measures implemented during a quantitative easing period, on the output; the other 47% were non-significant and a residual number, ~3%, were negative and significant. The average output increase after a QE related shock, at its maximum value, ranges between 0.35 and 0.4%. The length of the effect, during its statistically significant period, is in average ten to twelve months long. The literature selection here gathered identified eleven different transmission mechanisms from which the monetary based policy measures conducted by the BoJ, during the QE periods, conveyed its effect onto the Japanese output. Although distinguishable, these transmission channels are often related to one another, therefore, there hasn't been one pointed more often as the main driver of output growth. The small and short-lived positive effects of the QE policies are perceived as a combination of those transmissions mechanisms; these appear related with the BoJ's balance sheet and policy signaling. In the first case, the BoJ's large-scale asset purchase operations, intended to improve the financial market environment by taking over riskier assets and transferring more money into circulation, which in its turn, should allow banks to increase lending and ease credit, benefiting companies from both higher chances of investment and possible gains through the consequential improvement in the stock market environment. Second, the BoJ's commitment towards a sustained increase in inflation and active efforts, in order to boost the confidence of private agents, may have played a mild role on the output growth registered in these years.

In, Section 5, the funnel plots suggested that the effect sizes reported in published journals could be differently distributed from non-published papers, in terms of the relation between the magnitude of the effect size and their precision, making difficult to assess whether there is a true effect common to all literature. The same plots also suggested that there could be publication bias effect for both published and non-published papers, since the funnel shape, when visible, looked skewed to the right side of the plots. The results reported by the PET models, however, did not confirm in any case, the presence of publication bias, and also disproved the existence of a true effect regarding the effect of QE policy tools in the output; evidence that could be found for published studies but could not be extended when all studies were considered – published and non-published. Further attempts to screen for publication bias were introduced in the meta-probit analysis (Section 6), but none of them, confirmed it: published papers or papers with a higher impact ranking, do not have a higher probability to report positive and statistically significant effect sizes. Similarly, there was no discernable time trend, since the year of publication has been proven not significant. Furthermore, Section 6 presents evidence that certain elements found on the framework on the addressed literature may affect what is reported in terms of sign and significance. Those elements are related with the type of data used: the use of quarterly data and the increase of the number of observations make it more likely that papers report positive significant output growth in Japan. Inversely, if the analysis if solely focused on the first QE period, instead of focusing on a larger timeframe or other QE period – an element of a study's design – then the probability to report a positive and significant effect decreases. The other elements that seem to affect what is reported in the literature are related to the variables included in the models, from which results are withdrawn: choosing an industrial index as the output proxy or including a price level increases the chance of reporting a positive and significant output estimation. It decreases, if the model includes a bond yield or a monetary variable that represents either the money base or the money supply in the Japanese economy. The results regarding the choice of the 1st QE period, quarterly data, industrial output, bond yield and price level, are the strongest evidence found since they are heteroscedastic bias free.

8. Bibliography

Begg, C. B. & Berlin, J. A. 1988. Publication bias: a problem in interpreting medical data. ***Journal of the Royal Statistical Society***, (Series A) 151: 419–445.

Behrendt, S. 2013. ***Monetary Transmission via the Central Bank Balance Sheet***, Working Papers on Global Financial Markets, No. 49.

Berkmen, P. 2012. ***Bank of Japan's quantitative and credit easing: Are they now more effective?***. International Monetary Fund, IMF Working Papers: 12/2.

Cameron, A, C., & Douglas, L. M. 2015. A practitioner's guide to cluster-robust inference. ***The Journal of Human Resources***, 50(2): 317.

Davidson, R., & MacKinnon, J. G. 1984. Convenient Specification Tests for Logit and Probit Models. ***Journal of Econometrics***, *25*(3): 241–262.

De Michelis, A., & Iacoviello, M. 2016. ***Raising an inflation target: The Japanese experience with abenomics***. Board of Governors of the Federal Reserve System, International Finance Discussion Papers: 1168.

Dekle, R., & Hamada, K. 2015. Japanese monetary policy and international spillovers. ***Journal of International Money and Finance***, 52: 175-199.

Doucouliagos, H., & Stanley T, D. 2009. Publication selection bias in minimum-wage research? A meta-regression analysis. ***British Journal of Industrial Relations***, 47(2): 406-428.

Franta, M. 2011. ***Identification of monetary policy shocks in Japan using sign restrictions within the TVP-VAR framework***. St. Louis: Federal Reserve Bank of St Louis.

Fujiwara, I. 2006. Evaluating monetary policy when nominal interest rates are almost zero. ***Journal of the Japanese and International Economies***, 20(3): 434-453.

Gambacorta, L., Hofmann, B., & Peersman, G. 2012. ***The effectiveness of unconventional monetary policy at the zero lower bound: A cross-country analysis***. BIS Working Papers No 384.

Girardin, E., & Moussa, Z. 2011. Quantitative easing works: Lessons from the unique experience in Japan 2001-2006. ***Journal of International Financial Markets***, Institutions and Money, 21(4): 461-495.

Grauwe, P. D., & Storti, C. C. 2004. ***The effects of monetary policy: A meta-analysis***. CESifo Working Paper No. 1224.

Harada, Y., & Masujima, M. 2009. Effectiveness and transmission mechanisms of Japan's quantitative monetary easing policy. ***Japanese Economy***, 36(1): 48-105.
Harvey, A. C. 1976. Estimating Regression Models with Multiplicative Heteroscedasticity. ***Econometrica***, *44*(3): 461-465.

Havranek, T., & Rusnak, M. 2013. Transmission lags of monetary policy: A meta-analysis. ***International Journal of Central Banking***, 9(4): 39-75.

Hayashi, F. & Koeda, J. 2013. ***A Regime-Switching SVAR Analysis of Quantitative Easing***. CARF F-Series, CARF-F-322, Center for Advanced Research in Finance, Faculty of Economics, The University of Tokyo. Tokyo.

Honda, Y. 2014. The effectiveness of nontraditional monetary policy: The case of Japan. ***Japanese Economic Review***, 65(1): 1-23.

Honda, Y., Kuroki, Y., & Tachibana, M. 2013. An injection of base money at zero interest rates: Empirical evidence from the Japanese experience 2001-2006. ***Japanese Journal of Monetary and Financial Economics.*** 1(1): 1-24.

Hosmer, D. W., & Lemeshow, S. 2000. ***Applied Logistic Regression.*** 2nd ed. 392 p. Wiley Series in Probability and Statistics – Applied Probability and Statistics Section. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Inoue, T., & Okimoto, T. 2008. Were there structural breaks in the effects of Japanese monetary policy? Re-evaluating policy effects of the lost decade**.** ***Journal of the Japanese and International Economies***, 22(3): 320-342.

Iwata, S. 2010. Monetary policy and the term structure of interest rates when short-term rates are close to zero. ***Monetary and Economic Studies***, 28: 59-77.

Kagraoka, Y., & Moussa, Z. 2013. Quantitative easing, credibility and the time-varying dynamics of the term structure of interest rate in Japan. ***Journal of International Financial Markets, Institutions and Money***, 25: 181-201.

Kamada, K., & Sugo, T. 2006. ***Evaluating Japanese monetary policy under the non-negativity constraint on nominal short-term interest rates***. Bank of Japan Working Paper Series No. 06-E-17.

Kelley, K. & Preacher, J. K. 2012. On Effect Size. ***Psychological Methods***, 17(2): 137–152.

Kimura, T., & Nakajima, J. 2016. Identifying conventional and unconventional monetary policy shocks: A latent threshold approach. ***B.E.Journal of Macroeconomics***, 16(1): 277-300.

Kimura, T., Kobayashi, H., Muranaga, J., & Ugai, H. 2002. ***The effect of the increase in the monetary base on Japan's economy at zero interest rates***. IMES Disscusion Paper, Institute for Monetary and Economic Studies, Bank of Japan, 2002-NaN-22(19), 286.

Kluve, Jochen. 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.

Koetse, M. J., Groot, H. L. F. De, & Florax, R. J. G. M. (2009). A Meta-Analysis of the Investment-Uncertainty Relationship, ***Southern Economic Journal***, 76(1): 283–306.

Kurihara Y., & Fukushima A. 2015. Monetary Approach for Determining Exchange Rates and Recent Monetary Policy of Japan. ***International Journal of Financial Economics***, 4(1): 23-31.

Matsuki, T., Sugimoto, K., & Satoma, K. 2015. Effects of the bank of Japan's current quantitative and qualitative easing. ***Economics Letters***, 133: 112-116.

Mehrotra, A. 2009. The case for price level or inflation targeting – what happened to monetary policy effectiveness during the Japanese disinflation?. ***Japan and the World Economy***, 21(3): 280-291.

Mehrotra, A. N. 2007. Exchange and interest rate channels during a deflationary era –evidence from Japan, Hong Kong and china. ***Journal of Comparative Economics***, 35(1): 188-210.

Michaelis, H., & Watzka, S. 2014. ***Are there differences in the effectiveness of Quantitative Easing in Japan over time?***, MPRA Paper, No. 2014-35.

Morita, H. 2015. ***Japanese fiscal policy under zero lower bound of nominal interest rates: time-varying parameters vector autoregression***. Institute of Economic Research, Discussion Paper Series A, No. 627. Hitotsubashi University. Kunitachi.

Moussa, Z. 2010. ***The Japanese Quantitative Easing Policy under scrutiny: a time-varying parameter factor-augmented VAR model***. Munich Discussion Paper, No. 29429.

Nakajima, J. 2011a. Monetary policy transmission under zero interest rates: An extended time-varying parameter vector autoregression approach. ***B.E. Journal of Macroeconomics***: Topics in Macroeconomics, 11(1).

Nakajima, J. 2011b. Time-varying parameter VAR model with stochastic volatility: An overview of methodology and empirical applications. ***Monetary and Economic Studies***, 29: 107-142.

Nakajima, J., Kasuya, M., & Watanabe, T. 2011. Bayesian analysis of time-varying parameter vector autoregressive model for the Japanese economy and monetary policy. ***Journal of the Japanese and International Economies***, 25(3): 225-245.

Nakajima, J., Shiratsuka, S., & Teranishi, Y. 2010. ***The effects of monetary policy commitment: Evidence from time-varying parameter VAR analysis***. IMES Discussion Paper, Institute for Monetary and Economic Studies, Bank of Japan No. 2010-E-6.

Nakashima, K. 2011. ***Measuring Japanese monetary policy: an examination using the Bernanke-Mihov Econometric Methodology***. Konan Economic Papers V. 51, Number 1, 2, 3, 4: 47-79. Konan University. Higashinada-ku.

Neves, P., Silva, S., & Óscar Afonso. 2012. ***A meta-analytic assessment of the effects of inequality on growth***. CEF.UP Working Paper 2012-04.

Norton, E. C., Wang, H., & Ai, C. 2004. Computing Interaction Effects and Standard Errors in Logit and Probit Models. ***The Stata Journal***, *4*(2): 154–167.

Ohta, H. 2014. ***Evaluation of Monetary Easing Policy in Japan: Integration between the US and Japanese Markets***. Working Paper Series IR2014/3, 23. The International Studies Association of Ritsumeikan University. Kyoto.

Primiceri, G. E. 2005. Time varying structural vector autoregressions and monetary policy. ***Review of Economic Studies***, 72(3): 821-852.

Ridhwan, M. M., de Groot, H., L. F., & 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.

Rusnak, M., Havranek, T., & Horvath, R. 2013. How to solve the price puzzle? A meta-analysis. ***Journal of Money, Credit, and Banking***, 45(1): 37-70.

Saiki, A., & Frost, J. 2014. ***How does unconventional monetary policy affect inequality? Evidence from Japan***. Netherlands Central Bank, Research Department, DNB Working Papers.

Schenkelberg, H., & Watzka, S. 2013. Real effects of quantitative easing at the zero lower bound: Structural VAR-based evidence from Japan. ***Journal of International Money and Finance***, 33: 327–357.

Shirai, S. 2014. Japan's monetary policy in a challenging environment. ***Eurasian Economic Review***, 4(1): 3-24.

Stanley T, D. 2001. Wheat from chaff: Meta-analysis as quantitative literature review. ***The Journal of Economic Perspectives***, 15(3): 131-150.

Stanley, T. D. 2005. Beyond publication bias. ***Journal of Economic Surveys***, 19(3): 309-345.

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): 103-127.

StataCorp. 2013. *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

StataCorp. 2017. *Stata 15 Base Reference Manual*. College Station, TX: Stata Press.

Tachibana, M. 2015. ***Comparing the Macroeconomic Effects of Conventional and Unconventional Monetary Policy in Japan***. Discussion Paper New Series, 2015(3): 1-24. Osaka Prefecture University. Sakai.

Troug, H. A., & Murray, M. 2015. ***Quantitative Easing in Japan and the UK. An Econometric Evaluation of the Impacts of Unconventional Monetary Policy on the Returns of Aggregate Output and Price Levels***. MPRA Paper No. 68707.

Ueda, K. 2013. Response of asset prices to monetary policy under abenomics. ***Asian Economic Policy Review***, 8(2): 252-269.

Ugai, H. 2015. ***The transmission channels and welfare implications of unconventional monetary easing policy in Japan***. TCER Working Paper E-102. Hitotsubashi University. Kunitachi.

Usküla, L., & Pitzel, D. 2007. The Effect of Financial Depth on Monetary Transmission. ***The IUP Journal of Monetary Economics***. 5(2): 63-73.

9. Annex I

Table 9-1: Chronology of the Main Guidelines and Events that Characterize the Acting of Japanese Policy Makers in the Last 18 Years

February, 1999

- BoJ announces the first time establishment of the ZIRP regime, which ought to continue until an end to the deflationary scenario would be foreseen.

August, 2000

- The BoJ discontinued the ZIRP allowing the (uncollaterized) call rate to reach around 0.25 percent.

February, 2001

- The call rate is tightened from 0.25 to 0.125 percent.

March, 2001

- The BoJ introduces the first QE programme and re-establishes the ZIRP regime.

March, 2006

- Observing and forecasting favorable conditions for the Japanese economy, and predicting steady positive CPI inflation, the BoJ decided to switch targets from the current account balance to the call rate, thus, ceasing the QE programme.

July, 2006

- The ZIRP regime is discontinued, followed by a raise of the call rate to 0.25 percent.

August, 2006

- The change of the CPI base year from 2000 to 2005 transformed this month inflation from a positive value (February 2007 and October 2008 registered a core-CPI inflation of 0.5% under the 2000 base line), into a negative one.

October 2008

- Last month of the positive cycle of the core-CPI.

October, 2010

- Introduction of the Comprehensive Monetary Easing programme.

February, 2012

- Announcement of the continuation of the ZIRP and establishment of an Asset Purchase Program; both part of plan designated as "Enhancement of Monetary Easing", which would be in place until the purpose of pursuing a one percent CPI increase seemed secured.

January, 2013

- BoJ introduces inflation targeting.

April, 2013

- Announcement of the Quantitative and Qualitative Easing, aiming a year-on-year increase of the CPI until it was foreseeing a steady inflation around 2 percent within a period of two years.

April, 2014

- Consumers' tax was increased from 5 to 8 percent.

October, 2015

- BoJ re-estates the purpose of achieving the price stability target of two percent inflation rate.

June, 2016

- The increase of the consumer tax hike (from 8 to 10 percent) was announced to be postponed from April 2017 to October 2019.

Table 9-2: Descriptive Statistics

Variable	Description	Obs.	Mean	Median	Std. Dev.	Frequency (%)
Authorial Information						
Year of Publication	Latest publication.	104	2012.32	2013.00	2.99	2006 – 2.88 2007 – 7.69 2008 – 0.96 2009 – 12.50 2010 – 1.92 2011 – 13.46 2012 – 6.73 2013 – 5.77 2014 – 19.23 2015 – 13.46 2016 – 15.38

Type of Publication (1)	wp – working paper/mimeo; pp – published paper.	104	-	-	-	pp – 45.19 wp – 54.81
Type of Publication (2)	wp – working paper; gj – general journal; mj – monetary journal.	104	-	-	-	gj – 38.46 mj – 6.73 wp – 54.81
Are the Author(s) Associated with the Bank of Japan?	= 0 if no; = 1 if yes.	104	0.20	0.00	0.40	Yes – 20.19 No – 79.81

Data

QE Programs Comprehended in the Analysis' Timeframe	0 = if only the First QE Program is analyzed; 1 = if Other Timeframes (OTF) were used: - First QE Program and CME; QQE; First QE Program, CME (only); QQE (only) and CME and QQE;	104	-	-	-	First QE – 60.58 OTF – 39.42
Periodicity of the Time Series	Daily; Monthly; or Quarterly.	104	-	-	-	Daily – 1.92 Monthly – 53.85 Quarterly – 44.23

Number of Observations of the Analysis' Timeframe	L50 if lower than 50 obs. B50100 if between 50 and 100 obs. H100 if higher than 100 obs.	104	-	-	-	L50 – 12.50 B50100 – 43.27 H100 – 44.23
Mid-point of the Study's Timeframe	Converted in Quarterly data: YearQ1 – xxxx.25 YearQ2 – xxxx.5 YearQ3 – xxxx.75 YearQ4 – xxxx+1 1st line – Non-TVP (VAR-VEC and BVAR) models 2nd line TVP models 3rd line MSVAR models	104	2005.00 1999.47 1995.18	2004.00 1997.25 1994.50	5.22 3.51 4.59	Non-TVP – 62.51 TVP – 27.88 MSVAR – 9.61
Methodological Specifications						
Empirical Method	TVPVAR-BVAR-SwitchingVAR if - TVP-VAR; TVP-VAR with Stochastic Volatility; or TVP-FAVAR. - MSVAR; MS-FAVAR; or Regime	104	-	-	-	TVPVAR and Switching VAR – 37.50 VAR-VEC – 62.50

	Switching SVAR. - Bayesian SVAR VAR-VEC if - Recursive VAR; Recursive VAR with dummy; Non-linear VAR; Sign-restricted VAR; SVAR; or VAR. VEC					
Variable(s) that Measures the Output	GDP if - GDP; or Real GDP. GDP growth if - GDP growth; Real GDP growth. OG – Output gap (or proxy)if - GDP gap; Investment gap; Output gap. IO – Industrial Output if - Total Industry Activity Index; Industrial Production Index; Industrial Production; All Industry Activity Indices; Real	104	-	-	-	GDP growth and OG – 25.00 GDP – 12.50 IO – 61.54 UR – 0.96

	Production Index; Electric Power Consumption ; Economic Activity Factor. UR – Unemployment Rate.					
Other variables used in the regression						
Does the model have a Monetary Variable?	= 0 if no; = 1 if yes.	104	0.89	1.00	0.31	Yes – 89.42 No –10.58
Monetary variable	MB – Monetary Base if: - Monetary Base - Average Outstanding Account Balance - Current Account Balance - Reserves or Reserves Ratio MSt – Money Stock if: - Money Stock - Japanese Government	104	-	-	-	NA – 10.58 MB – 81.73 MSt – 7.69

	Bonds					
Does the model have a Second Monetary variable?	= 0 if no; = 1 if yes.	104	0.08	0.00	0.27	No – 92.31 Yes – 7.69
Second Monetary Variable	M2 (money stock) M3 (money stock)	104	-	-	-	NA – 92.31 M2 – 5.77 M3 – 1.92
Does the model have an Interest Rate of Reference?	= 0 if no; = 1 if yes.	104	0.54	1.00	0.50	No – 46.15 Yes – 53.85
Interest Rate of Reference	3MIR – 3-month interest rate CR – Call rate RR – Repo Rate	104	-	-	-	NA – 46.15 3MIR – 2.88 CR – 49.04 RR – 1.92
Does the model have a Price level (or Proxy)?	= 0 if no; = 1 if yes.	104	0.86	1.00	0.35	No – 14.42 Yes – 85.58
Price level (or Proxy)	CPI – CPI IR – Inflation Rate CCIG – Core CPI inflation Gap GDPD – GDP deflator	104	-	-	-	NA – 14.42 CCIG – 1.92 CPI – 57.69 GDPD – 0.96 IR – 25.00
Does the	= 0 if no; = 1 if yes.	10	0.27	0.00	0.4	No – 73.08

model have an Exchange Rate?		4			5	Yes – 26.92
Exchange Rate	NEER – Nominal Effective Exchange Rate NYDSR – Nominal Yen/Dollar Spot Rate REER – Real Effective Exchange Rate TWREFER – Trade Weighted Real Effective Foreign Exchange Rate	104	-	-	-	NA – 73.08 NEER – 7.69 NYDSR – 5.77 REER – 12.50 TWREFER – 0.96
Does the model have a Spread?	= 0 if no; = 1 if yes.	104	0.09	0.00	0.28	No – 91.35 Yes – 8.65
Spread	S5YCR – Difference between the 5-year JGB yield and the Call Rate S10YCR – Difference between the 10-year JGB yield and the Call Rate	104	-	-	-	NA – 91.35 S5YCR – 4.81 S10YCR – 3.85

Does the model have a Bond Yield?	= 0 if no; = 1 if yes.	104	0.18	0.00	0.39	No – 81.73 Yes – 18.27
Bond Yield	JGB – 10-year JGB yield or JGB yields	104	-	-	-	NA – 81.73 JGB – 18.27
Synthesized variables (= 1 if used; = 0 if not)	BOJPGB – Bank of Japan Purchase of Government Bonds	104	0.04	0.00	0.19	No – 96.15 Yes – 3.85
	BOJSP – Bank of Japan Stock Purchases	104	0.02	0.00	0.14	No – 98.08 Yes – 1.92
	SP – Stock Prices (or Stock Price Index)	104	0.24	0.00	0.43	No – 75.96 Yes – 24.04
Other Variables (= 1 if used; = 0 if not)	Average Lending Rate (on loans and discounts with maturity of less than one year at the time of origination)	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Bank of Japan ETFs Purchases	104	0.02	0.00	0.14	No – 98.08 Yes – 1.92
	Bank of Japan J – REITs Purchases	104	0.02	0.00	0.14	No – 98.08 Yes – 1.92
	Bank Lending in Japan	104	0.07	0.00	0.25	No – 93.27 Yes – 6.73
	Bank Share Prices	104	0.04	0.00	0.19	No – 96.15 Yes – 3.85
	CPI inflation of	10	0.02	0.00	0.1	No – 98.08

	Energy and Food (Exogenous Variable)	4			4	Yes – 1.92
	Commodity Price	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Condo Price Index	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Dummy Variables	104	0.03	0.00	0.17	No – 97.12 Yes – 2.88
	Gini Coefficient of Income Inequality	104	0.03	0.00	0.17	No – 97.12 Yes – 2.88
	Government Expenditure	104	0.09	0.00	0.28	No – 91.35 Yes – 8.65
	IOBOJMP – Indirect Observance of Bank of Japan Monetary Policy	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Interest Rate Factor (applicable to FAVAR models only)	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Japanese Exports	104	0.02	0.00	0.14	No – 98.08 Yes – 1.92
	Loans and Discounts in the Japanese Banking System	104	0.02	0.00	0.14	No – 98.08 Yes – 1.92
	Non-performing Loans in Japan	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Oil Inflation Rate	10	0.01	0.00	0.1	No – 99.04

		4			0	Yes – 0.96
	Price Level Factor (applicable to FAVAR models only)	104	0.02	0.00	0.14	No – 98.08 Yes – 1.92
	Yield Level Factor	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Yield Slope Factor	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Yield Curvature Factor	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
	Value of Civil Engineering Projects (government expenditure)	104	0.01	0.00	0.10	No – 99.04 Yes – 0.96
Observable Time Length of the IRF (in months)	L25 if lower than 25 months; B2648 if between 26 and 48 months; H48 if higher than 48 months. (Excludes TVP-VAR based estimates.)	104	-	-	-	L25 – 22.63 B2648 – 39.42 H48 – 13.87
Type of Shock (1)	OTS – Other Type of Shock SMB – Shock to the Monetary Base SMS – Shock to the Money Stock	104	-	-	-	OTS – 6.73 SMB – 52.88 SMS – 27.88 SSTIRR – 12.50

	SSTIRR – Shock to a Short-term Interest Rate of Reference					
Type of Shock (2)	QE; Non-QE; or Others.	104	-	-	-	QE – 80.77 Non-QE – 12.50 Others – 6.73
Confidence Intervals	68PCI if - 68% Confidence Intervals; - $\pm One$ - standard deviation Confidence Bands; - 16th and 84th Percentile Confidence Bands. 90PCI-1090PCB if - 90% Confidence Intervals - 10th and 90th Percentile Confidence Bands. 95PCI if - 95% Confidence Intervals; - $\pm Two$ - Standard	104	-	-	-	NA – 11.54 68PCI – 41.35 90PCI-1090PCB – 18.27 95PCI – 28.85

	deviation confidence bands.					
Output Variable in Levels or in First Differences?	0 = if Levels (LVS); 1= First Differences (FIRSTDIF)	104	0.38	0.00	0.49	LVS – 62.50 FIRSTDIF – 37.50
Monetary Policy Variable in Levels or in First Differences?	0 = if Levels (LVS); 1= First Differences (FIRSTDIF)	104	0.35	0	0.48	LVS – 65.38 FIRSTDIF – 34.62
Is the shock employed at a specific date?	= 0 if no; = 1 if yes.	104	0.29	0.00	0.46	No – 71.15 Yes – 28.85
The Date of the Beginning of the Shock (if applicable)	-	104	-	-	-	NA – 70.19 2002 – 1.92 2003 – 1.92 2004 – 2.88 2005 – 1.92 2006 – 6.80 2010 – 1.92 2011 – 1.92 2013 – 0.96 2014 – 0.96 2002:Q1 – 8.65 2003:Q3 – 0.96 2004:M02 – 1.92 2013:Q3 – 0.96

Estimates						
Accumulated Effect of the Shock's Impact on the Output Variable	= 0 if no; = 1 if yes.	104	0.11	0.00	0.31	No – 89.42 Yes – 10.58
Is the Effect of the Shock on the Output Statistically Valid? Thus verifying the Presence of Granger Causality (through the observation of the Confidence Intervals)?	= 0 if no; = 1 if yes.	104	0.48	-	0.50	No – 51.92 Yes – 48.08
Signal of the Shock's Impact in the Output Variable	1 if Significantly : - Pos. QE shock - Pos. Accommodative Non-QE shock - Negative Contractionary Non-QE shock 0 if Non-significant; -1 if Significantly: - Neg. QE shock	104	-	-	-	Overall Positive Effect – 50.00 Non-significant. – 47.12 Overall Negative Effect – 2.88

	- Neg. Accommodative Non-QE shock - Positive Contractionary Non-QE shock					
Persistence of the Shock's Impact in the Output Variable (in Months)	Intervals of two months ranging from *0 to 2* until *58 to 60* months.	104	-	-	-	NA – 60.58 0 to 2 – 4.81 6 to 8 – 6.73 8 to 10 – 10.58 10 to 12 – 3.85 12 to 14 – 2.88 16 to 80 – 0.96 18 to 20 – 1.92 24 to 26 – 0.96 26 to 28 – 1.92 28 to 30 – 0.96 30 to 32 – 0.96 32 to 34 – 0.96 38 to 40 – 0.96 58 to 60 – 0.96
Magnitude: Maximum Value of the Shock's Impact in the Output Variable (When Statistically Significant; in Intervals Percentage)	Intervals of 0.05 ranging from *[0 ; 0.05]* until *]3.95 ; 4]*. Registered as “NS” (non-significant) when the presence of Granger causality is not verified. Intervals registering null values omitted.	104	-	-	-	NS – 44.23 [-1.3; -1.25[– 0.96 [-0.7; -0.65[– 0.96 [-0.55; -0.5[– 0.96 [-0.35; -0.3[– 1.92] [-0.3; -0.25[– 1.92 [-0.05; 0[– 1.92

[0] – 1.92]0; 0.05] – 18.27]0.05; 0.1] – 0.96]0.1; 0.15] – 2.88]0.15; 0.2] – 0.96]0.2; 0.25] – 1.92]0.25; 0.3] – 0.96]0.3; 0.35] – 1.92]0.35; 0.4] – 0.96]0.4; 0.45] – 3.85]0.45; 0.5] – 0.96]0.55; 0.6] – 2.88]0.7; 0.75] – 0.96]0.75; 0.8] – 0.96]0.8; 0.85] – 0.96]1; 1.05] – 0.96]1.20; 1.25] – 0.96]1.25; 1.3] –

						0.96]1.65; 1.7] – 0.96]1.75; 1.8] – 0.96]3.95; 4] – 0.96
Transmission Channels Thought to Affect Output (1)	APP - Asset Purchase Program BBS – Bank Balance Sheets CEC – Credit Easing Channel DEIT – Direct Effect of Inflation Targeting IER – Increase in the Excess Reserves PRE – Portfolio Re-balancing Effect RIRC – Real Interest Rate Channel SPC – Stock Price Channel TCU – Transmission Channel Undefined TQE – Tobin's Q Effect WE – Wealth Effect	10 4	-	-	-	NA – 49.04 APP – 1.92 BBS – 0.96 CEC –0.96 DEIT – 0.96 IER – 8.65 PRE – 1.92 RIRC – 0.96 SPC – 5.77 TCU – 21.15 TQE – 2.88 WE – 4.81
Transmission Channel Thought to Affect the Output (2)	BBS-SPC – Bank Balance Sheets w/ Stock Price Channel FG-PRE – Forward	10 4	-	-	-	NA – 89.44 BBS-SPC – 2.88 FG-DEIT –

	Guidance w/ Portfolio Re-balancing Effect FG-DEIT – Forward Guidance w/ Direct Effect of Inflation Targeting WE-PRE – Wealth Effect w/ Portfolio Re-balancing Effect WE-TQE – Wealth Effect w/ Tobin's Q Effect	0.96 FG-PRE – 0.96 WE-PRE – 0.96 WE-TQE – 4.80

1) "NA" stands for "Not Available".
2) Transmission Channels Thought to Affect Output (1) identifies the transmission mechanisms pointed out in the literature when only one is mentioned.
3) Transmission Channels Thought to Affect Output (2) identifies the transmission mechanisms that were also pointed out in the literature but when another mechanism has been already mentioned.

Table 9-3: Descriptive Statistics for the Estimates – Persistence and Magnitude, Discriminating between QE and Non-QE shocks

	Obs.	Mean	Median	Maximum(2)	Minimum(2)	Std. Dev.	Jarque-Bera	Probability
Persistence – Non-QE(1)	8	4.75	3.50	11.00 (28 to 30)	1.00 (0 to 2)	3.81	0.0000	0.0000
Persistence – QE and Other(1)	33	4.82	3.00	15.00 (58 to 60)	1.00 (0 to 2)	3.92	1.1613	0.0000

Maximum Magnitude – Non-QE(1)	9	-7.89	-6.00	1.00]0; 0.05]	-26.00 [-1.3; -1.25[	8.51	1.5667	0.0000
Maximum Magnitude – QE and Other(1)	49	8.47	3.00	8.00]0.35; 0.4]	-7.00 [-0.35; -0.3[	1.37	4.5362	0.0000
1st Quarter – Non-QE	8	0.00	-1.00	1.00]0; 0.05]	-2.00 [-0.1; -0.05[	1.41	1.0433	0.0000
1st Quarter – QE and Other	56	1.98	1.00	48.00]2.35; 2.4]	-17.00 [-0.85; -0.8[	8.17	8.1505	0.0000
12th Month – Non-QE	4	-1.75	-2.00	1.00]0; 0.05]	-4.00 [-0.2; -0.15[	2.22	0.0000	0.0000
12th Month – QE and Other	47	4.74	1.00	44.00]2.15; 2.2]	-6.00 [-0.3; -0.25[	8.63	1.5938	0.0000
24th Month – Non-QE	4	-3.75	-3.50	-1.00 [-0.1; -0.05[	-7.00 [-0.35; -0.3[	3.20	0.0000	0.0000
24th Month – QE and Other	47	5.30	1.00	41.00]2; 2.05]	-9.00 [-0.45; -0.4[	1.07	5.9300	0.0000
36th Month – Non-QE	3	-3.33	-4.00	-1.00 [-0.1; -0.05[	-5.00 [-0.25; -0.2[	2.08	0.0000	0.0000
36th Month – QE and Other	23	1.48	1.00	7.00]0.30; 0.35]	-5.00 [-0.25; -0.2[	3.84	1.2892	0.0000
48th Month – Non-QE	3	-2.33	-3.00	-1.00 [-0.1; -0.05[	-3.00 [-0.3; -0.25[	1.15	0.0000	0.0000
48th Month – QE and Other	23	0.00	1.00	6.00]0.25; 0.3]	-6.00 [-0.3; -0.25[	2.89	1.0416	0.0000

1) Only for these categories the results reported concern statistically significant observations.
2) Persistence in months; Magnitude in output percentage intervals.
3) Except for Persistence and Maximum Magnitude (QE/Other and Non QE), the observations do not account for the TVP-Switching category (Bayesian SVAR one Obs accounted for).

Table 9-4: Percentage Distribution of the Variations in the VAR Family Identified in the Literature Selection (104 Obs.)

▪ TVPVAR and Switching VAR	%	▪ VAR	%

TVP-VAR	25.96	VAR	35.58
TVP-FAVAR	1.92	Structural VAR	8.65
MSVAR	5.77	Vector Error Corrected (VEC)	0.96
MS-FAVAR	1.92	Recursive VAR	5.77
Regime-switching SVAR	1.92	Recursive VAR with dummy	2.88
		Signed-restricted VAR	5.44
		Non-linear VAR	0.96
▪ Bayesian VAR	0.96		

Table 9-5: Explicit Reference to Transmission Mechanisms

Transmission Channel or effect	Nr. Of Studies
Asset Purchase Program	1
Bank Balance Sheets	2
Credit Easing Channel	1
Direct Effect of Inflation Targeting	1
Forward Guidance	2
Increase in the Excess Reserves	2
Portfolio Re-balancing Effect	1
Real Interest Rate Channel	1
Stock Price Channel	1
Tobin's Q Effect	2
Transmission Channel Undefined	12
Wealth Effect	2

Figure 9-a: Effect Sizes by Year of Publication/Release (104 Obs.)

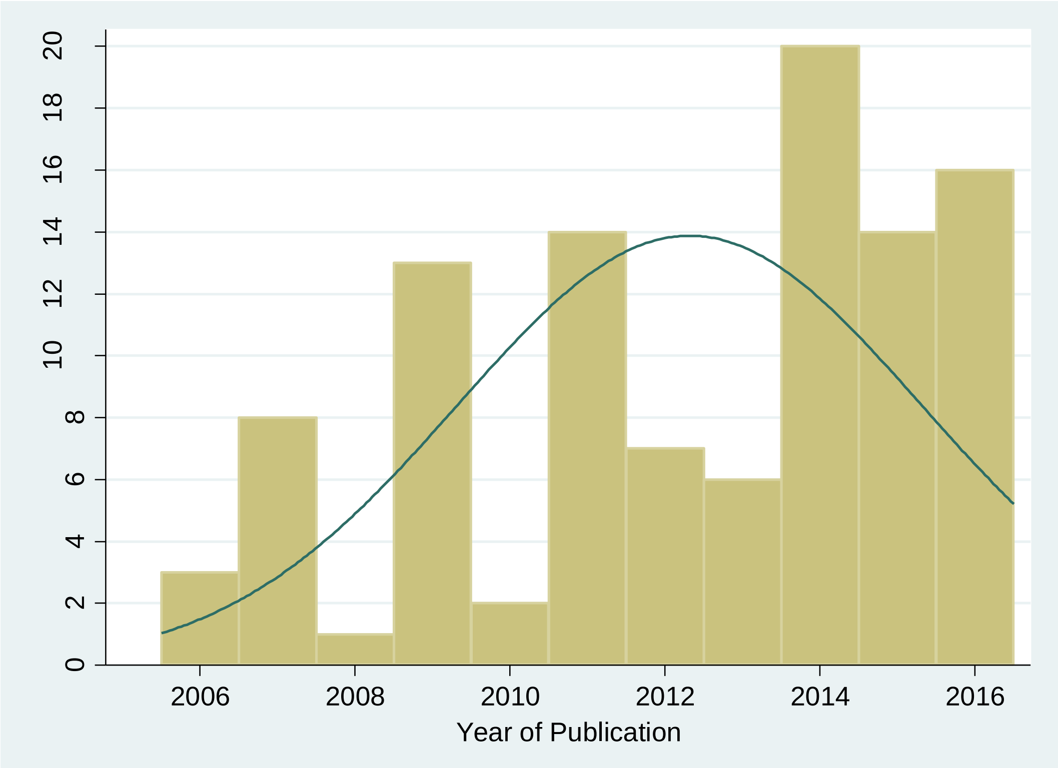
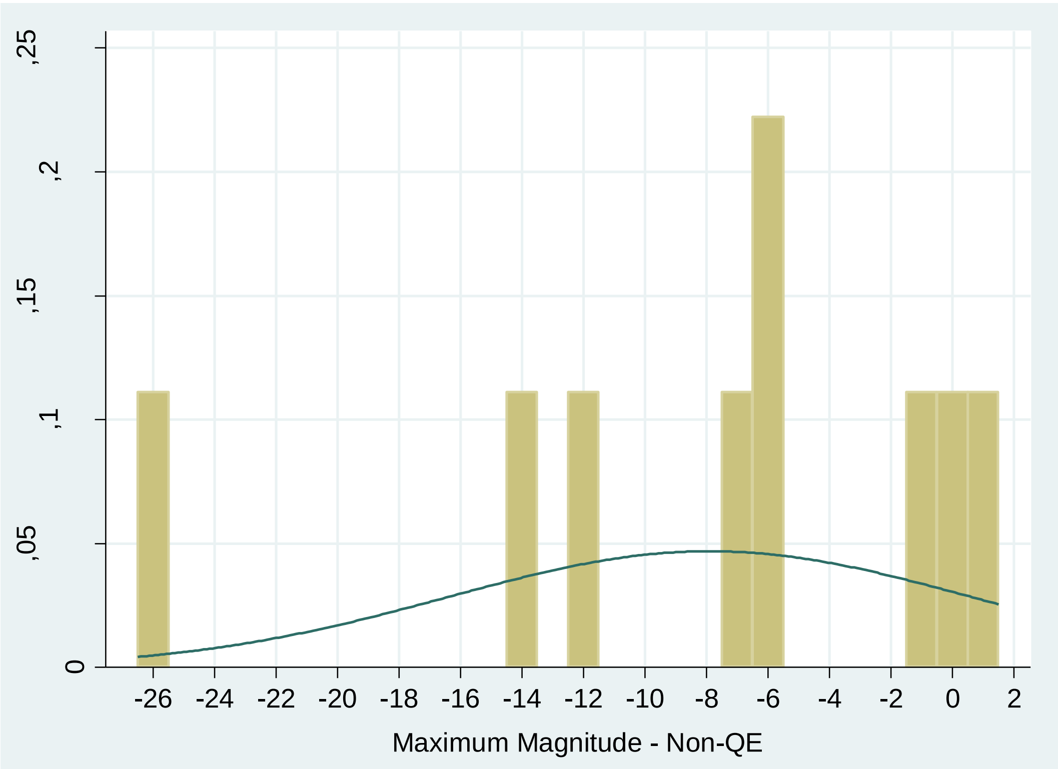


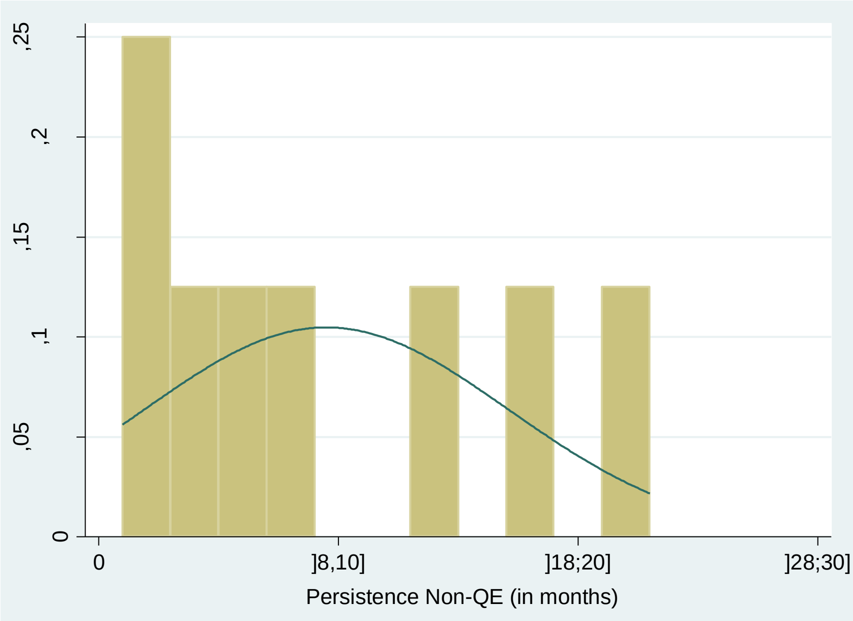
Figure 9-b: Density of Intervals of Maximum Magnitude of the Output Response to a Non-QE Shock



1) The first ten (negative) intervals are read as follows (in percentage): 1 –]0 - 0,05]; 2 –]0,05 - 0,1]; 3 –]0,1 - 0,15]; 4 –]0,15 - 0,2]; 5 –]0,2 - 0,25]; 6 –]0,25 - 0,3]; 7 –]0,3 - 0,35]; 8 –]0,35 - 0,4]; 9 –]0,4 - 0,45]; 10 –]0,45 - 0,5].

2) Intervals correspond to the maximum value attained by the output response to a MP shock during a statistically significant period (9 Obs.)

Figure 9-c: Persistence of the Shock's Impact in the Output Variable during a Statically Significant Period (in Months; 8 Obs.)



10. Annex II

Table 10-1: PET Results for Maximum Magnitude Effect-Sizes

	MLME – All studies	Cluster OLS – All Studies	MLME – Published Studies	Cluster OLS – Published Studies	MLME w/ Dummy	Cluster OLS w/ Dummy
Observations	48	48	26	26	48	48
Number of Clusters	18	18	9	9	18	18
β_1 (Intercept)	-21.608 (69.127)	2.348 (69.270)	-8.858 (61.560)	17.822 (55.738)	-4.508 (65.830)	9.916 (71.688)
$\beta_0(\sqrt{n_{ij}})$	8.325 (5.446)	6.599 (6.194)	4.841 (4.531)	2.463 (4.306)	8.045 (5.137)	7.147 (6.350)

$d: BoJ = 1$; otherwise $= 0$	-	-	-	-	-82.747 (55.223)	-80.863 *** (25.636)
Wald test (p-value)	0.1264	-	0.2853	-	0.0928	-
F test (p-value)	-	0.3016	-	0.5830	-	0.0180 **
R^2	-	0.0650	-	0.0466	-	0.1665
Jarque-Bera test (p-value)	0.0000	0.0005	0.0217	0.1623	0.0000	0.0008
LR test (p-value)	0.0000 ***	-	0.0000 ***	-	0.0000 ***	-
Intra-class Correlation Ratio	0.763	-	0.912	-	0.737	-

Notes: parameter's standard errors reposted in parenthesis (robust for Cluster OLS). Model with Dummy distinguishes between effect sizes reported by authors associated with the Bank of Japan (*=1*) and those that don't (*=0*). *,**,*** if statistically significant at 10%, 5% and 1% confidence level, respectively. The null hypothesis of the LR test states that MLME model is equal to a simple linear model. The Wald test for MLME models is used instead of the F test, and its null hypothesis states that the coefficients estimated are both simultaneously equal to zero. ICC ratio ranges from [0;1]. 0 stands for no correlation within clusters (no variance to explain at cluster level) and 1 stands for maximum correlation (no variance at the individual level, i.e. the observations within clusters have the same value). We reject the null hypothesis of normally distributed errors whenever Jarque-Bera's test *p*-value is less than 1%.

Table 10-2: PET Results for 1st Quarter Effect-Sizes

	MLME – All studies	Cluster OLS – All Studies	OLS – All Studies (1)	MLME – Published Studies	Cluster OLS – Published Studies	MLME w/ Dummy	Cluster OLS w/ Dummy	OLS w/ Dummy
Observations	55	55	55	25	25	55	55	55
Number of Clusters	15	15	-	7	7	15	15	-

β_1 (Intercept)	-16.759 (29.195)	-16.759 (32.298)	-16.759 (29.741)	152.878 (106.237)	81.687 (88.806)	-18.601 (30.046)	-18.601 (35.934)	-18.601 (30.901)
$\beta_0(\sqrt{n_{ij}})$	3.215 (3.162)	3.215 (4.394)	3.215 (3.221)	-16.943* (10.137)	-8.649 (11.392)	3.460 (3.301)	3.460 (4.889)	3.460 (3.395)
$d: BoJ = 1;$ otherwise $= 0$	-	-	-	-	-	-17.385 (67.703)	-17.385 (36.285)	17.385 (69.629)
Wald test (p-value)	0.3093	-	-	0.0947*	-	0.5768	-	-
F test (p-value)	-	0.4765	0.3228	-	0.4765	-	NA	0.5974
R^2	-	0.0184	0.0184	-	0.1404	-	0.0196	0.0196
Jarque-Bera test (p-value)	0.0001	0.0001	0.0001	0.0189	0.0000	0.0001	0.0001	0.0001
LR test (p-value)	1.0000	-	-	0.0000***	-	1.0000	-	-
Intra-class Correlation Ratio	0.000	-	-	0.977	-	0.000	-	-

Notes: parameter's standard errors reposted in parenthesis (robust for Cluster OLS). Model with Dummy distinguishes between effect sizes reported by authors associated with the Bank of Japan (*=1*) and those that don't (*=0*). *,**,*** if statistically significant at 10%, 5% and 1% confidence level, respectively. The null hypothesis of the LR test states that MLME model is equal to a simple linear model. The Wald test for MLME models is used instead of the F test, and its null hypothesis states that the coefficients estimated are both simultaneously equal to zero. ICC ratio ranges from [0;1]. 0 stands for no correlation within clusters (no variance to explain at cluster level) and 1 stands for maximum correlation (no variance at the individual level, i.e. the observations within clusters have the same value). We reject the null hypothesis of normally distributed errors whenever Jarque-Bera's test p-value is less than 1%.

1) We cannot overrule heteroscedasticity in the errors of this model according to a Breusch-Pagan Het. Test (*p-value* < 0.01).

Table 10-3: PET Results for 12th Month Effect-Sizes

	MLME – All studies	Cluster OLS – All Studies	MLME – Published Studies	Cluster OLS – Published Studies	MLME w/ Dummy	Cluster OLS w/ Dummy
Observations	46	46	25	25	46	46
Number of Clusters	13	13	7	7	13	13
β_1 (Intercept)	32.589 (45.290)	62.844 (57.700)	76.685 (54.313)	63.754 62.770	29.718 (46.659)	62.565 (60.892)
$\beta_0(\sqrt{n_{ij}})$	-0.199 (4.333)	-3.142 (5.483)	-7.271 (5.205)	-4.931 (6.704)	0.179 (4.582)	-3.106 (5.953)
$d: BoJ = 1;$ otherwise $= 0$	-	-	-	-	-17.714 (70.495)	-2.500 (32.424)
Wald test (*p*-value)	0.9632	-	0.1624	-	0.9680	-
F test (*p*-value)	-	0.5772	-	0.4897	-	NA
R^2	-	0.0194	-	0.0933	-	0.0194
Jarque-Bera test (*p*-value)	0.0000	0.0000	0.1953	0.5493	0.0000	0.0000
LR test (*p*-value)	0.0042 ***	-	0.0000 ***	-	0.0040 ***	-
Intra-class Correlation Ratio	0.363	-	0.865	-	0.364	-

Notes: parameter's standard errors reposted in parenthesis (robust for Cluster OLS). Model with Dummy distinguishes between effect sizes reported by authors associated with the Bank of Japan (=*1*) and those that don't (=*0*). *,**,*** if statistically significant at 10%, 5% and 1% confidence level, respectively. The null hypothesis of the LR test states that MLME model is equal to a simple linear model. The Wald test for MLME models is used instead of the F test, and its null hypothesis states that the coefficients estimated are both simultaneously equal to zero. ICC ratio ranges from [0;1]. 0 stands for no correlation within clusters (no variance to explain at cluster level) and 1 stands for maximum correlation (no variance at the individual level, i.e. the observations within clusters have the same value). We reject the null hypothesis of normally distributed errors whenever Jarque-Bera's test *p*-value is less than 1%.

Table 10-4: PET Results for 24th Month Effect-Sizes

	MLME – All studies	Cluster OLS – All Studies	MLME – Published Studies	Cluster OLS – Published Studies	MLME w/ Dummy	Cluster OLS w/ Dummy
Number of Clusters	13	13	7	7	13	13
Observations	46	46	25	25	46	46
β_1 (Intercept)	11.378 (54.842)	51.679 (79.404)	-64.118 (59.316)	-20.513 (38.468)	4.815 (56.238)	48.811 (81.598)
$\beta_0(\sqrt{n_{ij}})$	2.553 (5.164)	-0.956 (6.232)	7.646 (5.681)	3.516 (4.844)	3.436 (5.445)	-0.582 (6.491)
$d{:}\, BoJ = 1$; otherwise $= 0$	-	-	-	-	-40.463 (81.072)	-25.659 (19.736)
Wald test (*p*-value)	0.6210	-	0.1783	-	0.7810	-
F test (*p*-value)	-	0.8806	-	0.4952	-	NA
R^2	-	0.0012	-	0.0740	-	0.0033
Jarque-Bera test (*p*-value)	0.0000	0.0001	0.0258	0.0016	0.0000	0.0001
LR test (*p*-value)	0.0000 ***	-	0.0003 ***	-	0.0000 ***	-
Intra-class	0.490	-	0.879	-	0.489	-

Correlation Ratio						

Notes: parameter's standard errors reposted in parenthesis (robust for Cluster OLS). Model with Dummy distinguishes between effect sizes reported by authors associated with the Bank of Japan (*=1*) and those that don't (*=0*). *,**,*** if statistically significant at 10%, 5% and 1% confidence level, respectively. The null hypothesis of the LR test states that MLME model is equal to a simple linear model. The Wald test for MLME models is used instead of the F test, and its null hypothesis states that the coefficients estimated are both simultaneously equal to zero. ICC ratio ranges from [0;1]. 0 stands for no correlation within clusters (no variance to explain at cluster level) and 1 stands for maximum correlation (no variance at the individual level, i.e. the observations within clusters have the same value). We reject the null hypothesis of normally distributed errors whenever Jarque-Bera's test p-value is less than 1%.

11. Annex III

Table 11-1: Correlation Matrix of the Variables used in the Univariate Analysis

(obs=104)

	year	qe	period~y	io	mv_1	mv_2	price	ci_1
year	1.0000							
qe	0.7486	1.0000						
periodicity	-0.3268	-0.1928	1.0000					
io	-0.4002	-0.2520	0.7285	1.0000				
mv_1	0.2204	0.2286	-0.0202	0.0866	1.0000			
mv_2	-0.2879	-0.2329	0.0391	0.0799	-0.6106	1.0000		
price	-0.3431	-0.2849	0.1855	0.1818	-0.0524	-0.0869	1.0000	
ci_1	0.4615	0.4414	-0.3138	-0.2192	0.2959	-0.2424	-0.1555	1.0000
ci_2	-0.5859	-0.4855	0.1037	0.0335	-0.2018	0.3058	0.2778	-0.7925
obs	-0.0014	0.0382	0.2052	0.0942	-0.1106	0.2394	0.0760	0.3979
log_obs	0.0633	0.0787	0.1010	0.0647	-0.0809	0.2005	0.0224	0.4870
int_obs	0.0036	0.0438	0.1808	0.0710	-0.1206	0.2505	0.0880	0.3936
bond_yield	0.2421	0.2296	0.0703	0.0157	-0.1628	0.0503	0.1941	0.0073
shock_a	-0.0975	-0.0663	0.3230	0.2041	0.5009	-0.3058	0.3253	-0.0290
shock_b	0.2429	0.2443	-0.2233	-0.2136	-0.0944	0.3838	-0.5381	0.0004
shock_c	-0.1528	-0.1939	-0.1549	-0.0154	-0.5270	-0.0493	0.2003	0.0362

	ci_2	obs	log_obs	int_obs	bond_y~d	shock_a	shock_b	shock_c
ci_2	1.0000							
obs	-0.2036	1.0000						
log_obs	-0.2505	0.9461	1.0000					
int_obs	-0.1872	0.9977	0.9478	1.0000				
bond_yield	-0.2967	-0.0326	-0.0183	-0.0230	1.0000			
shock_a	0.1191	-0.1546	-0.1286	-0.1463	0.0475	1.0000		
shock_b	-0.0285	0.0921	0.0393	0.0860	-0.0720	-0.6588	1.0000	
shock_c	-0.1184	0.0911	0.1182	0.0874	0.0219	-0.5170	-0.3034	1.0000

1) Relevant correlations over 0.7 (in absolute terms) marked in red.

Table 11-2: Heteroscedastic Probit Model Estimation Results

	Model 4-a	Model 4-b	Model 5
QE	-131.215 (1819.485)	-0.535 (0.753)	-
Periodicity	-	0.494 (0.683)	-
IO	635.233 (9999.188)	-	5.048 (4.700)
Price level	660.492 (10265.43)	0.975 (0.671)	-
Bond Yield	-1226.405 (17837.44)	-1.043 (1.957)	-8.596 (7.450)
MV1	-	-	-10.812 (9.192)
MV2	-	-	1555.701 (106021.7)
Shock a	-	-	7.197 (6.517)
Shock b	-	-	-
Shock c	-	-	
Obs	-	-	0.050 (0.046)
Int_obs	-	-	-
Year	-	-	-

Constant	-635.233(9999.187)	-0.722 (0.649)	1.118 (2.828)
Log-likelihood	-52.87363	-53.93275	-40.59066
Nr. of Iterations to achieve Convergence	36	8	18
"Likelihood-ratio Test of lnsigma^2 = 0"	$Chi^2(4) = 4.67$ (0.3225)	$Chi^2(4) = 1.64$ (0.8021)	$Chi^2(6) = 17.95$ (0.0064)
"lnsigma^2" Z-stat.'s *p*-values (below)	Model 4-a	Model 4-b	Model 5
QE	0.683	0.343	-
Periodicity	-	0.751	-
IO	0.823	-	0.009
Price level	0.749	0.883	
Bond Yield	0.997	0.578	0.060
MV1	-	-	0.052
MV2	-	-	0.889
Shock a	-	-	0.007
Obs	-	-	0.011
Note 1	-	-	Same results when Int. Obs instead of Obs

1) Standard Errors in parenthesis.

2) "Likelihood-ratio Test of lnsigma^2 = 0" is a Stata test that checks the overall significance of the regressors within the probit's variance, that are associated to the index variables. A test's *p*-value under 5% suggests that the heteroscedastic probit model is less biased, in regards to the parameters, than the homoscedastic standard probit model.

3) "lnsigma^2" Z-stat.'s *p*-value regard the statistical significance reported by Stata, of the regressors within the probit's variance, that are associated to the index variables.

Figures Model 4-a:

Figure 11-a: Interaction QE*Bond, Significance and Corrected Effect Respectively.

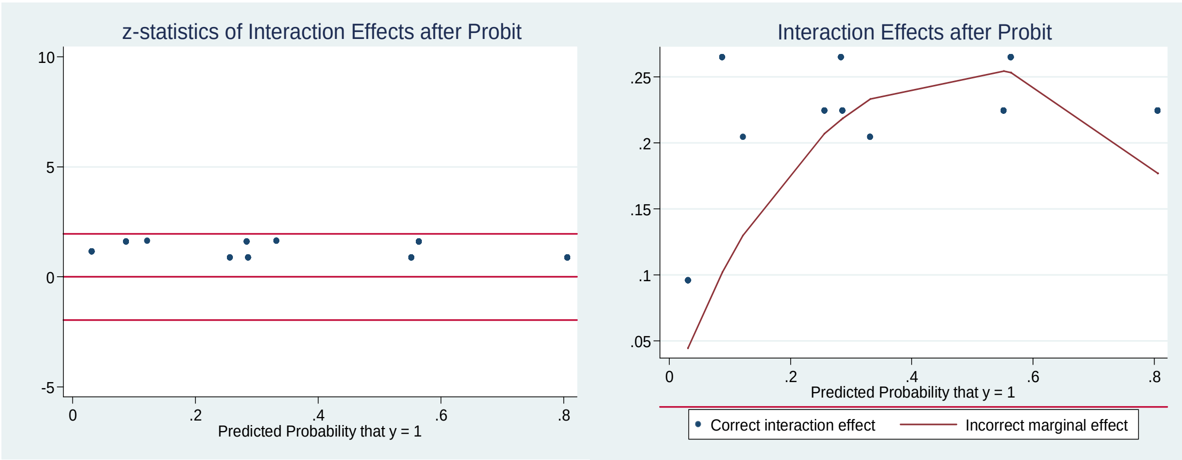


Figure 11-b: Interaction IO*Bond, Significance and Corrected Effect Respectively.

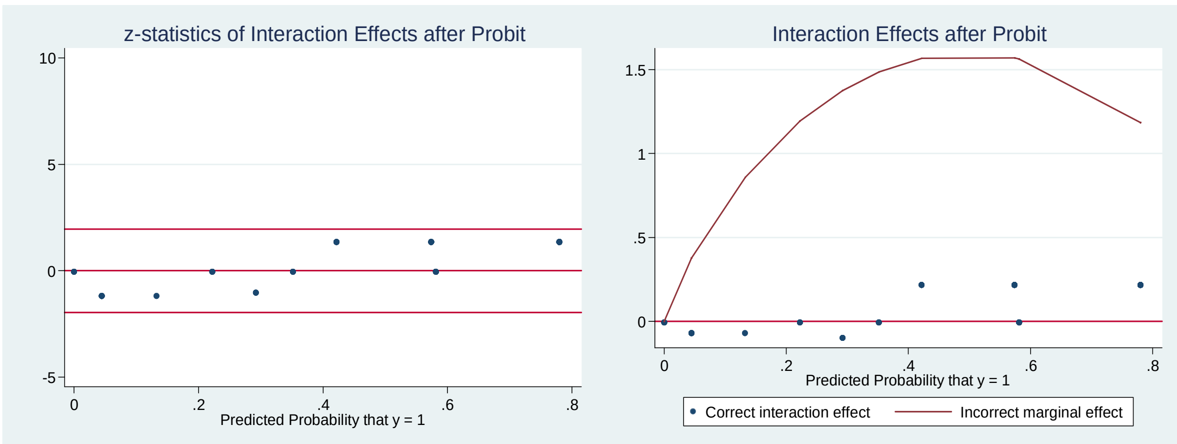
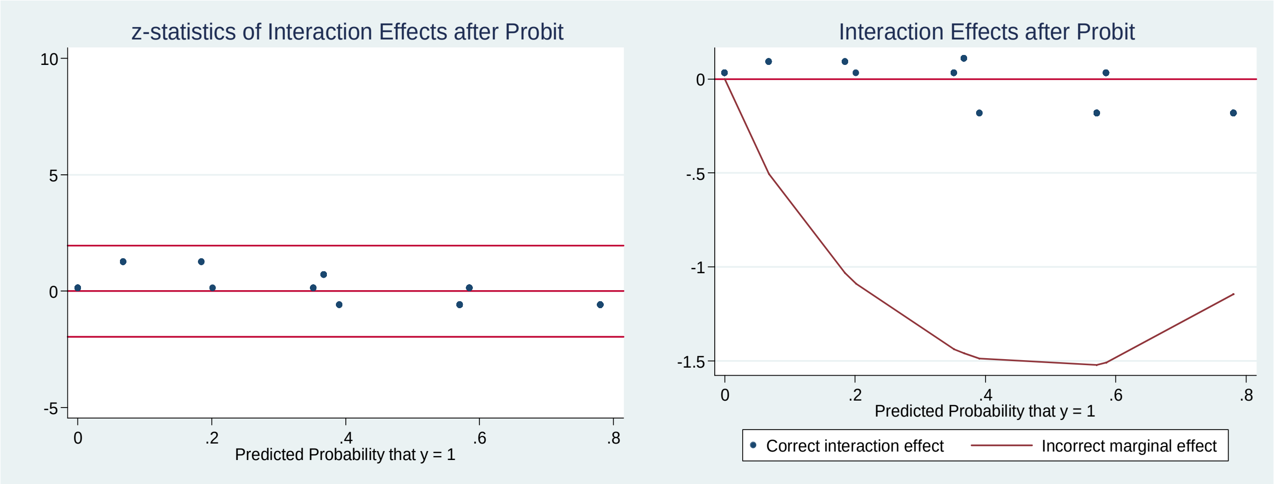


Figure 11-c: Interaction IO*Price, Significance and Corrected Effect Respectively.



Figures Mode 4-b:

Figure 11-d: Interaction QE*Bond, Significance and Corrected Effect Respectively.