The effects of weather on wheat yields in Victoria, Australia: an empirical study

Jess Arkesden ..................................................................................................................................................Page 4

The impact of quantitative easing on financial markets in the United Kingdom

James Barrie ..................................................................................................................................................Page 41

The Impact of the ‘One-Child Policy’ on China’s Aggregate Household Savings – An Econometric Analysis

Lele Ding ......................................................................................................................................................Page 77

How Has Inequality Been a Cause of Violence in Post-dictatorship Brazil?

Samson Hart ..................................................................................................................................................Page 111

Is a trade-off between justice and efficiency inevitable? A comparative study of European welfare states between 2001 and 2013

Kasia Lipka ...................................................................................................................................................Page 141

Prison Labour: Should the UK follow the US’s example?

Jake Lush McCrum .......................................................................................................................................Page 171

How are ratings agencies perceived following the US financial crisis? What were their main shortfalls? Insights from academics and professionals

Timothy Sapsford ........................................................................................................................................Page 198

An econometric study into the impact of the current account on house price growth in the United States

George Vernon ..............................................................................................................................................Page 230
The Effect of the UK’s Smoke Free Law on Smoking Prevalence, Tobacco Related Hospital Admissions and NHS Expenditure

*Kate Wales*...Page 260

The Impacts of Immigration on Wages and Unemployment in England: An Empirical Investigation

*George Wibberley*...Page 291

The Impacts of Digital Currency on China’s Monetary System

*Qianru Xiang*...Page 321
The effects of weather on wheat yields in Victoria, Australia: an empirical study.

Jess Arkesden

This study looks at the impacts that weather variables have on wheat yields in the state of Victoria, Australia. Both basic measures (temperature and precipitation) and complex measures (evapotranspiration rate and the Standardised Precipitation Index) are used. Econometric analysis is used to estimate the effect these variables have on wheat yields for 1953-2011, whilst including non-linear, interaction terms and dummy variables for extreme weather events. The regression results show that variations from the average temperature and rainfall have a negative impact on yields. However, drought has a larger and more significant impact on wheat in Victoria than changes in precipitation.
1. Introduction

Evidence suggests that weather patterns are changing. It is widely accepted that this will have an impact on crop production. Australia, renowned for its climate, has undoubtedly begun to experience this change. Since the colonial era, Australia has been a producer of wheat for both trade and consumption (GrainGrowers Ltd, 2011). Output has grown significantly, with wheat now the most valuable crop under production in the country (Australian Bureau of Statistics (ABS), 2007). The future, however, is a cause for concern and is largely dependent on Climate Change. This study aims to identify the extent to which temperature and precipitation changes from 1953 to 2011 have impacted wheat yields in the state of Victoria.

This study begins by explaining the background and context of the wheat industry in Australia, and specifically Victoria. Wheat is most successful in the Southwest and Southeast of Australia (PricewaterhouseCooper (PwC), 2011b). Victoria ranks third for yields in the country (ABS, 2007), highlighting the importance of the state’s production.

Weather variables are summarised. Precipitation has fluctuated over the study period with some cyclical patterns present but displaying an overall declining trend (Bureau of Meteorology (BoM), 2015). Preliminary data from Swan Hill (a farm in the Mallee region) shows some correlation between decreasing rainfall and yields (van Rees et al., 2011). Similarly, temperatures appear to be increasing. In the past 100 years, the temperature in Australia has increased by 0.7°C (Department of Agriculture, Fisheries and Forestry (DAFF), 2006). This has coincided with growing concern over Climate Change globally, coinciding with the increased frequency and intensity of droughts (Wang et al., 1992).

A review of the current body of literature is then conducted; there is, however, limited research on Victoria specifically. The key study conducted by van Rees et al. (2011) focuses on the Mallee and Wimmera regions only. Hence, a wider body of literature is covered here analysing other states in Australia to gauge the extent to which – and how – weather variables are impacting wheat yields. Overall, yields are negatively affected by rising temperatures and falling rainfall, regardless of the state (Anwar et al., 2007; Ludwig and Asseng, 2006; van Rees et al., 2011).

Climate Change has been identified as a key threat to the future (Anwar et al., 2007; Department of Environment and Primary Industries (DEPI), 2014). Whilst this dissertation does not seek to predict future changes and impacts, it is worthwhile acknowledging the current stance of key writers in the field. The Victorian government is working to develop new strains of grain that can withstand rising
temperatures and have reduced water demand (DEPI, 2014). Research by Anwar et al. (2007) prior to this found that yields are falling based on a variety of Climate Change Models. Ultimately, the literature demonstrates uncertainty in the future as a direct result of weather changes. This is subsequently supported by past data both within the current body of literature and the results of this study’s empirical analysis.

Hence, the subsequent Section details the empirical Model for this study, the methodology, results and discussion. The Model used is based on Blanc’s (2012) work. Although this focused on sub-Saharan Africa and a variety of crops, the study specifies two Models – one using basic weather indicators (precipitation and temperature); the other using complex measures (Standardised Precipitation Index (SPI) and Evapotranspiration Rate (ET)). The latter measures are previously unused in analysis of Australian crop production. Hence, using Blanc (2012) as a basis of this study, a new insight will be given into the effects of weather on crop yields in Victoria.

Data is collected from both ABS (2013) and BoM (2015) databases, with SPI and ET calculated from this. The results highlight non-linear relationships between weather variables and wheat yields, supporting the results of the current literature. Interestingly, a lack of rainfall has the largest impact on crops, but temperature changes present greater cause for concern given predictions of future changes. Fluctuations in rainfall have occurred throughout time, but a 3°C rise in temperature as suggested by some Models (Anwar et al., 2007) would have catastrophic impacts on wheat yields.

The conclusion of this dissertation is therefore that yields have been negatively impacted by weather pattern changes, making the future of the Victorian wheat industry increasingly uncertain.

2. Background

2.1 The Wheat Industry

Wheat is one of the major grain crops produced globally. In the past 40 years, global consumption has doubled (PwC, 2011a). Accounting for only 3% of global output, Australia is not a major producer (ABS, 2007). However, the country equates to 15% of annual world trade (ABS, 2007) with key markets including Indonesia, Malaysia and Japan (GrainGrowers Ltd, 2011). The value of this trade is merely 4% of Australian GDP, but totals to 30% of Australia’s exports (Luo et al., 2003). This makes wheat Australia’s most valuable and largest crop under production (ABS, 2007).
Australia’s history of wheat production dates back to the colonial era. From a meagre 5.4 tonnes harvested in 1790 (ABS, 2007), to 21,834 kilotons in 2009-2010 (PwC, 2011a), the wheat industry has grown substantially. Figure 1, taken from the GrainGrowers Ltd (2011) report on the wheat industry, shows that the boom in production occurred from the 1950s onwards, but has shown signs of slower growth in recent years as determined by the trend line.

**Figure 1: Australian Wheat Production, 1860-2010**

In early years, poor rainfall and inappropriate wheat varieties plagued production (ABS, 2007). Advancements in machinery and transportation allowed for larger areas to be cultivated and for wheat to be traded more freely (ABS, 2007). However, it was wheat breeding that enabled output to increase dramatically.

The value of this production is subject to fluctuations. War, depression and volatile market forces all affect prices. The European Union, Eastern Europe and the United States are key competitors (Linehan et al., 2012). In 2003-2004, the harvest had a gross value of A$5.6 billion, with exports at over A$3.4 billion (ABS, 2007). Global imports of wheat are projected to grow even further, with estimates being 135% greater in 2050 than 2007 (Linehan et al., 2012). This would equate to 2% annual growth of Australia’s exports.
Wheat is grown across Australia but is most successful in the wheat belt of the Southeast and Southwest where rainfall is between 230mm and 500mm (PwC, 2011b). The East coast of Australia centres on domestic production. Approximately 5 million tonnes remains in Australia each year, half for production and half for animal feed (ABS, 2007).

Out of the eight states, Victoria ranks fourth for area under wheat production and total production, and ranks third for yield – ahead of New South Wales, Queensland and South Australia, with yields above the countrywide average (ABS, 2007). Appendix 1 contains graphs illustrating the change in area and production for the state for the time period concerned in this study (1953-2011). Figure 2 uses this information to demonstrate how yields have changed – this is the dependant variable used for the empirical analysis. Production fluctuates greatly in the state, but the general trend is for rising output and yields (as shown by the trend line in Figure 2). Wimmera and Mallee are two regions that are noticeably more productive than the countrywide average (GrainGrowers Ltd, 2011). These are situated in the Northwest of the state.

**Figure 2: Wheat Yields for the State of Victoria, 1953-2011**

![Wheat Yields for the State of Victoria, 1953-2011](image)

Source(s): Author’s calculations; Data Source: ABS (2013)

Victoria is, however, experiencing a decrease in Water Use Efficiency (WUE) that is leading to reductions in yields for harvest cycles with lower rainfall (GrainGrowers Ltd, 2011). The 1990s was the warmest decade on record, with further temperature increases predicted (Bindi and Howden, 2004). This will occur simultaneously with more variable rainfall making the future for Victorian wheat increasingly uncertain.
2.2 The Weather

Weather patterns have a profound impact on wheat yields, with Climate Change a major concern for Australia’s future. Climate Change is therefore one of the key challenges for the wheat industry (ABS, 2007). Global warming affects wheat in three ways: the increase in CO₂ concentration; rising temperatures; and changes to rainfall and subsequent evaporative demands of the soil (Wang et al., 1992). In the last 100 years, temperatures across Australia have risen 0.7°C on average and days of extreme heat have become more frequent (DAFF, 2006). Rainfall has simultaneously come to be more unpredictable and intense (Bindi and Howden, 2004).

This study will focus on rainfall and temperature to determine how these have affected wheat yields in Victoria from 1953 to 2011. Precipitation has fluctuated quite dramatically in the study period, with clear drought and flood years present. The harvests of 1982-1983, 1994-1995 and 2002-2003 coincided with substantial droughts (DAFF, 2005). The effect of this is reflected in the low productivity and value of the crop in those years.

Figure 3 illustrates annual precipitation in Victoria. Years of drought appear to be becoming more frequent since the 1990s. Interestingly, this coincides with the aforementioned record high temperatures. However, 2010 was one of the wettest years on record, ending a 13-year drought period that was categorised by higher minimum and maximum temperatures (van Rees et al., 2011).

Figure 3: Annual Precipitation for Victoria (mm), 1953-2011

Source(s): Author’s calculations; Data Source: (BoM, 2015)
Drought periods can therefore be short-term or long-term. Short-term droughts are single years in which rainfall is significantly lower than average, reducing crop yields (van Rees et al., 2011). Long-term droughts consist of any length of time where rainfall is below average, with temperatures often simultaneously higher than usual (van Rees et al., 2011). Figure 4, taken from van Rees et al. (2011), highlights three single year droughts (starred events) and three multi-year droughts (red line events) for Swan Hill. This is a farm in the Mallee region of Victoria and so is an interesting parallel to the empirical Section of this study. As can be observed, whilst not entirely predictable, there does appear to be a somewhat cyclical pattern to rainfall.

**Figure 4: Historical Rainfall, Swan Hill, Mallee Region, 1891-2010**

![Rainfall graph]

Source(s): van Rees et al. (2011)

Similar patterns can be observed for temperature, with a general upward trend. Figure 5 illustrates the average annual temperature for Victoria across the wheat growing regions. The 13-year drought period is reflected in above average temperatures, with another noticeable peak in 1982 (preceding the aforementioned low yields of the 1982-1983 crop). Interestingly, the temperature in 1992 was below average whilst the rainfall was above average and the crop yield was poor. This highlights the complexity of weather and its effects on yields, demonstrating that achieving optimal crop yields is highly uncertain. Any combination of high or low rainfall and temperature can have an impact.

Predictions for the future vary. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) believes that average temperatures in Victoria could increase somewhere between 1°C and 3°C by 2030, whilst rainfall increases up to 20% during summer months and decreases up to 10% in winter months (Wang et al., 1992). Yields are optimal when rainfall aligns with the start of the growing season and it is drier at the end of the season to increase the protein content of the grain (PwC, 2011b). The growing season runs from April through to October (autumn through to spring). Hence the predicted changes would forecast a fall in yields.
Figure 5: Average Annual Temperature for Victoria (°C), 1953-2011

Source(s): Author’s calculations; Data Source: BoM (2015)

This Section has highlighted the significance of the wheat industry (globally, within Australia, and for Victoria as a state). Data shows that yields have increased on average in Victoria but have recently faced vast fluctuations. Weather is known to be a contributing factor. Basic weather data demonstrates further fluctuations that, preliminarily, appear to coincide with changes in yield.

Consequently, the aim of this study is to assess the effects of weather variables on wheat yields in Victoria. The Models specified have not yet been applied to research on the state. However, a number of other studies need to be understood to highlight the research so far. These have been conducted for both Australia and other wheat producing nations to determine how weather patterns have affected wheat yields. Some have then been combined with Climate Change predictions to estimate future yields. The subsequent Section will assess this body of literature.

3. Literature Review

The combined effects and interaction of all aspects of Climate Change determine the net impact on wheat yields (Ludwig and Asseng, 2006). Studies focusing specifically on the impact of weather on wheat production in Victoria are scarce. There is, however, a significant body of literature for other states in Australia. This Section will discuss the relevant findings looking first at results for temperature changes and then rainfall. Climate Change scenarios will then be covered to explain the existing approaches to the future. There are a number of limitations to the current research that will subsequently be briefly discussed.
3.1 Temperature

A number of studies have considered temperature and its impact on yields. Wheat yields are seriously affected by days with temperatures in excess of 30°C or below 2°C during September to November (van Rees \textit{et al}., 2011). Since 1900, the average temperature in Australia rose between 0.7°C (DAFF, 2006) and 0.8°C (Anwar \textit{et al}., 2007). CSIRO (2001 cited in DAFF, 2006) estimates a further rise of between 0.4°C and 2.0°C by 2030. However, there is a debate between whether the consequences of rising temperatures are positive or negative for wheat production.

Temperature increases affect wheat both directly and indirectly (Ludwig and Asseng, 2006). Heat stress is a direct cause of crop failure and reduced yields (van Herwaarden, 1998 cited in Ludwig and Asseng, 2006). The phonological phase – time for capture and use of light and water – is decreased, stunting growth (Anwar \textit{et al}., 2007). Rising water demand is an indirect consequence (Lawlow and Mitchell, 2000; cited in Ludwig and Asseng, 2006). This can lead to increases in the evapotranspiration rate (Wang \textit{et al}., 1992). This is the rate at which water evaporates from soil and transpires from plants (Blanc, 2012). As noted in the previous Section, decreased water use efficiency (WUE) is a growing concern in Victoria. Subsequently, evapotranspiration rates will be tested further in this study.

Location and soil type are consequently found to be two key aspects in the effects of temperature change. Different soils result in varied yield changes when tested under higher temperatures (Ludwig and Asseng, 2006). Sandy soil was found to be at maximum yields and so any further global warming will reduce yields. A similar study on clay soil tested under Mediterranean conditions (van Ittersum \textit{et al}., 2003) found that temperature increases up to 3°C had a positive effect on yields. At higher temperatures yields decreased.

However, tests in Ludwig and Asseng’s (2006) study have only been conducted for Western Australia. Results found that northern areas responded negatively to temperature increases whilst in southern areas yields rose (Ludwig and Asseng, 2006). In general however, temperature increases lead to a fall in yields as the growing season is reduced in length (Ludwig and Asseng, 2006). Similarly, Luo \textit{et al}., (2005) found that high rainfall areas in South Australia were affected more by temperature changes.

Given the shift in growing season (ABS, 2007), and the classification as a higher rainfall state (Wang \textit{et al}., 1992), this pattern may be present in Victoria. This study will analyse this further as derived in the subsequent Section.
Individual farm-level analysis in van Rees *et al.* (2011) did make some progress regarding temperature change in Victoria. The Mallee and Wimmera regions were the focal points. Trends since 2005 show that the Mallee region has experienced more days above 30°C than the long-term average (van Rees *et al.*, 2011). Simultaneously, yields decreased an average of 1.1 tonnes per hectare. Wimmera is also trending towards more days of extreme heat. In six of the eight years from 2002 to 2010, days of extreme heat were again above the long-term average and coincided with the end of the growing season (van Rees *et al.*, 2011). This study aims to expand on this research by analysing a larger data set for Victoria as a whole and for a longer period of time.

### 3.2 Rainfall

There has been a significant body of literature focusing on rainfall, particularly in Australia. Lack of moisture is a major constraint across the country (Colls, 1993). Hence precipitation patterns are a focal point for the future of the grains industry in general, not just wheat.

Wheat areas are determined by soil type and rainfall (ABS, 2007). Rain should fall mostly in winter and spring at a minimal annual rate of 400-600mm. Yields decrease when crop available moisture (CAM) falls (GrainGrowers Ltd, 2011). Victoria has experienced significant CAM decline of 1.3mm per year since 1990 (GrainGrowers Ltd, 2011). Most is attributed to declining winter rainfall. The only way to maintain yields, according to industry experts, is to increase the WUE at the same rate as CAM declines through crop breeding (GrainGrowers Ltd, 2011). In Victoria, WUE has also declined.

Contrastingly to the results for temperature changes, clay soil is affected more by rainfall changes than sandy soils. This is because clay soil stores water closer to the surface and so evapotranspiration occurs quicker, hence less rainfall increases the risk of drought dramatically (Ludwig and Asseng, 2006). Regardless of type, soil water prior to anthesis (when the crop is fully functioning) must be sufficient to ensure maximum crop yields (Wang *et al.*, 1992). Even after anthesis, sufficient CAM is critical to achieving full yield potential. Hence optimal rainfall is central to successful wheat production.

In semi-arid regions, an increase in rainfall may increase output (Ludwig and Asseng, 2006). However, a 25% fall in precipitation across the year could decrease yields by more than 25% (van Ittersum *et al.*, 2003). Contrastingly, high rainfall zones suffer from nutrition loss and are at risk of floods when rainfall increases (Ludwig and Asseng, 2006). Both negatively impact yields.
Changes in rainfall can, however, have both positive and negative effects. According to both Luo *et al.* (2005) and Ludwig and Asseng’s (2006) results, the effects of changing rainfall are not linear. Increases and decreases of significant levels will reduce yields. Hence, rainfall plotted against yield will be an inverted-U shape curve. Luo *et al.* (2005) note this is an area in need of further research. Consequently, this will be explored further in this study.

Interestingly, Ludwig and Asseng (2006) found that a 15% decrease in rainfall could be compensated for by a 2°C rise in temperature. However, this is only for northern and central areas of Western Australia covered in the study. Contrarily, Luo *et al.* (2005) found that – for all locations in the study – median grain yield is positively correlated with rising rainfall when combined with increasing CO$_2$ concentration but negatively with rising temperatures. Rainfall, however, was the main determinant of changing yields with a coefficient of 0.88 compared to temperature (0.055) and CO$_2$ (0.0089) (Luo *et al.*, 2005).

Again, there is limited research on Victoria. Ludwig and Asseng (2006) concluded that decreasing precipitation would be the main threat to Western Australian wheat crops. Winter rainfall has declined between 10% and 20% since 1970 (Ludwig and Asseng, 2006). Comparably, in South Australia (the driest Australian state) wetter sites were more sensitive to Climate Changes than rising CO$_2$ (Luo *et al.*, 2003). Hence even smaller changes in rainfall may impact crop yields here more significantly.

The available research on the Mallee and Wimmera regions reflect most of these results. Conclusions of the research support that wheat yields are affected greatest by growing season rainfall (GSR, April to October) (van Rees *et al.*, 2011). The correlation between falling GSR and yields is extremely strong. This will be assessed further in the subsequent Section for the whole of Victoria, as compared to just the Mallee and Wimmera regions.

For these areas though the data is compelling. The average GSR in southern parts of the Mallee region decreased from 239mm in 1990-2001 to 156mm in 2002-2009 (van Rees *et al.*, 2011). If GSR falls below 150mm wheat will struggle to grow at all. The three-year running averages for GSR against yields showed strong negative correlations (van Rees *et al.*, 2011). Similarly, the GSR in the Wimmera region has decreased since the 1990s with yields dictated by these changes. The harvests of 2002, 2004 and 2006 experienced GSR of 89mm, 65mm and 126mm less than average respectively (van Rees *et al.*, 2011). These years all experienced below average yields.
A study on Birchip, a farm in the Mallee region, found that annual rainfall was extremely variable. The general trend was for lower rainfall in hotter decades (Anwar et al., 2007). This highlights the link between weather variables. Interdependence of variables will be explored further in the subsequent Section. For now, Anwar et al. (2007) concluded that the estimated decrease of 7% in rainfall in the next 30 years will have a significant impact when combined with other variables, but alone it will not cause widespread crop failure.

3.3 Climate Change

As has been established, Climate Change poses the greatest threat to the future of Australian wheat production. It is therefore valuable to summarise some of the literature on the future of the industry and of Climate Change scenarios.

Climate Change is a public good, meaning everyone is responsible for it and experiences the effects (Naughten, 1993). Therefore, it is widely considered that there are two ways to tackle Climate Change. The first is mitigation – offset global warming by, for example, reducing emissions (Naughten, 1993). This needs to be achieved on an international scale to be effective. The second is adaptation – reduce the damage experienced from Climate Change, which can be done on a regional scale (Naughten, 1993).

The Victorian government has subsequently made large investments into the future of the wheat industry through knowledge acquisition. The Grains Research and Development Corporation and the Victorian government have each invested A$3million in the construction of an Australian Grains Genebank at Horsham, north-western Victoria (DEPI, 2014). Here they acquire, test, develop and distribute genetic resources to create wheat strains that will adapt to Climate Change. However, so far it has been found that yields can be increased but the grain quality is falling (DEPI, 2014). This will have the largest impact on bread production, which requires higher quality grain.

Anwar et al. (2007) based their Model analysis on four climate variables (rainfall, maximum and minimum temperature, solar radiation) to calculate per degree warming on a monthly basis. This produced low to high global warming scenarios. Future wheat yields are highest under the low global warming Model as is to be expected, but coincides with increases in CO$_2$ concentration (Anwar et al., 2007). Yields fell by 29%, but with enhanced CO$_2$ this was 25% (Anwar et al., 2007). Similarly, McKeon et al. (1988; cited in Luo et al., 2003) found that Climate Change was more of an issue than rising CO$_2$ in a study of Queensland. Rising CO$_2$ levels are not exponentially beneficial though, and cannot offset
the negative impacts of increasing temperatures and decreasing rainfall. All of the projected climate scenarios decrease yields (Anwar et al., 2007).

Based on the performance of the wheat industry over the past 100 years, Victorian farmers should be able to overcome issues Climate Change presents. However, adaptation will be key to this success and will ultimately reach an upper bound if the extent of Climate Change exceeds predictions (ABS, 2007). The Carbon Farming Initiative will ideally induce farmers into sustainable practices by offering compensation packages (PwC, 2011a). Much is still to be done to secure the future of the industry.

3.4 Opportunities in the Current Body of Literature

As has been explained in this Section, there has been limited research into the effects of weather variables on Victorian wheat specifically. Evidence from other studies suggests that falling rainfall and rising temperatures have had, and will continue to have, a negative impact on crop yields. Additional factors, such as adaptation and the role of CO$_2$ concentrations, alter the degree to which yields change but ultimately they are falling.

Hence, the subsequent Section will use Blanc’s (2012) work as a basis for the empirical Model. Whilst this is a study of sub-Saharan Africa, the Model takes into account basic and advanced weather variables as will be explained. The literature reviewed in this Section will be taken into consideration throughout the analysis, including data sources and current findings for other states. This study seeks to build on this body of literature by analysing a larger dataset for Victoria than has been previously used. This includes a longer period of time, larger area (not just the Mallee and Wimmera regions) and with more weather variables included in the Models.

4. Empirical

4.1 Variable and data description

4.1.1 Yield and Area

Basic economic theory offers a production function in which output is dependent on land, labour and capital inputs. Wheat yields (measured in tonnes per hectare) will be the dependent variable in this Model, as marginal productivity of the land is key to maximising output efficiently. The area of land cultivated will be included as an independent variable to demonstrate how productivity decreases as marginal land is cultivated. Output and area under wheat production between 1953 and 2011 for the state of Victoria are obtained from the Australian Bureau of Statistics (2013) historical database. Yield is then calculated.
4.1.2 Precipitation

Yields are affected by a number of factors. The purpose of this study is to assess the effects of weather. The two main variables are therefore precipitation and temperature. Average precipitation during the growing season is calculated as this most affects yields (van Rees et al., 2011). Growing season rainfall (GSR) occurs between April and October.

Observations of average monthly rainfall were obtained for 30 weather stations using the Bureau of Meteorology’s (BoM, 2015) weather database. Specific weather stations were chosen to cover all wheat producing areas across Victoria as rainfall varies greatly across the state. An average for the GSR was then calculated to represent the state as a whole.

4.1.3 Standardised Precipitation Index

A more complex rainfall measure – the Standardised Precipitation Index (SPI) – is also calculated to take account of extreme precipitation events. The SPI takes long-term, monthly precipitation data and calculates an index for 3-, 6-, 12-, 24- or 48-month timescales (World Meteorological Organisation (WMO), 2012). This study selects the 6-month timescale for April to September as this best reflects the GSR.

The SPI compares the precipitation for April-September against the same period for all chosen years. The data is fitted to a probability distribution, transformed to a normal distribution, and so presents the mean SPI as zero (Edwards and McKee, 1997 cited in WMO, 2012). Positive values demonstrate a greater than median rainfall and negative values illustrate a less than median rainfall. The critical values are presented in Table 1.

Table 1: SPI Values

<table>
<thead>
<tr>
<th>SPI values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0 and more</td>
<td>Extremely wet</td>
</tr>
<tr>
<td>1.5 to 1.99</td>
<td>Very wet</td>
</tr>
<tr>
<td>1.0 to 1.49</td>
<td>Moderately wet</td>
</tr>
<tr>
<td>-0.99 to 0.99</td>
<td>Near normal</td>
</tr>
<tr>
<td>-1.0 to -1.49</td>
<td>Moderately dry</td>
</tr>
<tr>
<td>-1.5 to -1.99</td>
<td>Severely dry</td>
</tr>
<tr>
<td>-2.0 and less</td>
<td>Extremely dry</td>
</tr>
</tbody>
</table>

Source(s): WMO (2012)
Dummy variables for droughts and floods are used to take into account extreme weather events. Droughts were given a dummy value of 1 if the SPI was less than -1.0 and floods were given a dummy value of 1 if the SPI was greater than 1.0.

4.1.4 Temperature
Abnormally high or low temperatures can negatively impact crops. The temperature variable uses the average maximum and minimum temperature for each month from the selected 30 weather stations to give an average monthly temperature. This is then transformed into an annual average temperature for 1953 to 2011.

4.1.5 Evapotranspiration Rate
The additional complex measure used here is the evapotranspiration (ET) rate. This combines water loss from soils through evaporation and from crops via transpiration (Blanc, 2012). If water needs are not met, the crop yields can be affected. Due to data limitations, the reference ET rate (ETo) is calculated using the Hargraeves equation (Blanc, 2012):

\[
ETo = 0.0023(T_{avg} + 17.8)(T_{max} - T_{min})^{0.5}R_a
\]

Average temperature \(T_{avg}\), average maximum temperature \(T_{max}\) and average minimum temperature \(T_{min}\) are calculated from annual data as explained above. The \(R_a\) term is solar exposure (radiation), obtained from the BoM’s database. It should be noted, \(R_a\) data is only available for each weather station from 1993 to 2011 and so an average for this period is calculated and used for all ETo rates generated for 1953 to 2011. Hence the ETo equation can be transformed into Equation 2 where \(\overline{R_a}\) is the mean radiation for the chosen time period.

\[
ETo = 0.0023(T_{avg} + 17.8)(T_{max} - T_{min})^{0.5}\overline{R_a}
\]

4.1.6 Omitted variables
Several variables are excluded from this study due to data limitations. Fertiliser type and usage, crop variety selection, crop management, diseases, pests and soil quality may all vary with weather patterns. There is insufficient data (or access) to use these variables. The focus is therefore solely on the impacts of weather on yields.
4.2 Production functions

Using the variables defined above, two Models are specified for this study (as derived in Blanc (2012)). These are shown in Table 2. Firstly, the temperature and precipitation (T-P) Model uses the basic weather variables as summarised in Equation 3. Equation 4 depicts the more complex variables in the ETo and SPI (ET-SPI) Model, which includes drought and flood dummies. Both of these Models use interaction terms to account for effects between variables (Blanc, 2012).

<table>
<thead>
<tr>
<th>Model Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Equation 3: T-P Model</strong></td>
</tr>
<tr>
<td>( Y_t = f(A_t, T_t, T_t^2, P_t, P_t^2, T \times P_t, T_t^2 \times P_t, T_t \times P_t^2) )</td>
</tr>
<tr>
<td>Where at time ( t ), ( Y ) is the yield, ( A ) the area cultivated, ( T ) temperature and ( P ) precipitation.</td>
</tr>
<tr>
<td><strong>Equation 4: ET-SPI Model</strong></td>
</tr>
<tr>
<td>( Y_t = f(A_t, ETo_t, ETo_t^2, SPI_t, Drought_t, Flood_t, ETo \times SPI_t, ETo^2 \times SPI_t, ETo \times Drought_t, ETo \times Flood_t) )</td>
</tr>
</tbody>
</table>

Source(s): Blanc (2012)

4.3 Methodology:

Equations 3 and 4 are then transformed into regressions (equations 5 and 6). Blanc (2012) uses log-transformed values of yield and area to estimate elasticities, whilst the other variables are not log-transformed to produce semi-elasticities. This allows the assessment of, for example, a 1 °C temperature change or 10mm precipitation change on yields. This study calculates the change in each variable (first difference) – specified in StataCorp (2013) as a variable minus its value in the previous time period using: \( changeInX = X - L.X \). This allows us to analyse the trends in effects of the variables as they change over time.

Table 3 contains the Ordinary Least Squares (OLS) regressions, showing the fully specified equations for each Model.

Descriptive statistics for key variables in both regressions are listed in Table 4. Table 4 gives a maximum yield between 1953 and 2011 that is over 8 times larger than the minimum. This highlights that yield can differ greatly through time. The standard deviation, meanwhile, shows that the yield varies by 0.51 tonnes per hectare (tonnes/Ha) from the average value (almost 31%). The aim of this
study, therefore, is that the independent explanatory variables chosen may be able to explain some of this change.

Table 3: Model Regression Equations

<table>
<thead>
<tr>
<th>Model Regression Equations</th>
<th>Equation 5: T-P Model Regression Equation</th>
<th>Equation 6: ET-SPI Model Regression Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[ \Delta \ln Y = \beta_1 \Delta \ln A + \beta_2 \Delta T + \beta_3 \Delta T^2 ]</td>
<td>[ \Delta \ln Y = \beta_1 \Delta \ln A + \beta_2 \Delta ETo + \beta_3 \Delta ETo^2 ]</td>
</tr>
<tr>
<td></td>
<td>[ + \beta_4 \Delta P + \beta_5 P^2 + \beta_6 T. P + ]</td>
<td>[ + \beta_4 \Delta SPI + \beta_5 \Delta Drought + \beta_6 \Delta Flood ]</td>
</tr>
<tr>
<td></td>
<td>[ \beta_7 T^2. P + \beta_8 T. P^2 + \beta_9 ]</td>
<td>[ + \beta_7 \Delta ETo. SPI + \beta_8 \Delta ETo^2. SPI ]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ + \beta_9 \Delta ETo. Drought + \beta_{10} \Delta ETo. Flood + \beta_{11} ]</td>
</tr>
</tbody>
</table>

The descriptive statistics for SPI in Table 4 also show that droughts and floods occur given that the maximum and minimum are greater than 1.0 and -1.0 respectively as defined previously. Extreme weather events included in the regression may demonstrate some significance in their role. Contrastingly, temperature appears to have a very small spread of 11.09°C to 13.99°C giving a standard deviation of 0.63°C from the mean. However, given that Climate Change scenarios highlight the significance of a 2°C rise (Luo et al., 2005) these temperature variations may have an impact on yields in the sample used in this study.
Table 4: Descriptive Statistics for Key Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (tonnes/Ha)</td>
<td>59</td>
<td>1.65</td>
<td>0.51</td>
<td>0.32</td>
<td>2.70</td>
</tr>
<tr>
<td>Area (Ha)</td>
<td>59</td>
<td>1,160,944</td>
<td>278,773</td>
<td>633,400</td>
<td>1,801,100</td>
</tr>
<tr>
<td>Precipitation (mm)</td>
<td>59</td>
<td>360.66</td>
<td>92.72</td>
<td>165.38</td>
<td>571.71</td>
</tr>
<tr>
<td>SPI (standard deviation from median)</td>
<td>59</td>
<td>0.0037</td>
<td>1.01</td>
<td>-2.51</td>
<td>2.05</td>
</tr>
<tr>
<td>Drought (dummy)</td>
<td>59</td>
<td>0.15</td>
<td>0.36</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Flood (dummy)</td>
<td>59</td>
<td>0.19</td>
<td>0.39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>59</td>
<td>12.48</td>
<td>0.63</td>
<td>11.09</td>
<td>13.99</td>
</tr>
<tr>
<td>ETo (mm/day)</td>
<td>59</td>
<td>1.70</td>
<td>0.08</td>
<td>1.53</td>
<td>1.89</td>
</tr>
</tbody>
</table>

4.4 Results and discussion

4.4.1 T-P Model regression

Appendix 2 contains the regression results for a variety of Model specifications (1-4). The analysis will focus on the results presented in Table 5. This compares Model 3 against Model 4 to ensure the best analysis is provided. Model 3 has the best fit ($R^2$) whilst Model 4 has the greatest number of statistically significant variables.
### Table 5: T-P Model Regression Results

| Model 3 – dependant variable $\Delta \ln Y^1$ | Coefficient | Robust Std. Err. | $t^3$ | $P > |t|^4$ |
|--------------------------------------------|-------------|-----------------|-------|-----------|
| $\Delta \ln A$                             | 0.7762      | 0.3852          | 2.02  | 0.049     |
| $\Delta T$                                 | -5.0313     | 19.0237         | -0.26 | 0.793     |
| $\Delta T^2$                               | 0.1780      | 0.6936          | 0.26  | 0.798     |
| $\Delta P$                                 | -0.0085     | 0.3759          | -0.02 | 0.982     |
| $\Delta P^2$                               | -0.0000     | 0.0002          | -0.22 | 0.829     |
| $\Delta T.P$                               | -0.0020     | 0.0517          | -0.04 | 0.969     |
| $\Delta T^2.P$                             | 0.0001      | 0.0017          | 0.06  | 0.954     |
| $\Delta T.P^2$                             | 0.0000      | 0.0000          | 0.36  | 0.718     |
| Constant                                   | -0.0002     | 0.0571          | -0.00 | 0.997     |

| Model 4 – dependant variable $\Delta \ln Y^5$ | Coefficient | Robust Std. Err. | $t^6$ | $P > |t|$ |
|---------------------------------------------|-------------|-----------------|-------|--------|
| $\Delta \ln A$                             | 0.8381      | 0.3934          | 2.13  | 0.038  |
| $\Delta T^2$                               | 0.0145      | 0.0054          | 2.72  | 0.009  |
| $\Delta P^2$                               | 0.0000      | 0.0000          | 1.86  | 0.069  |
| $\Delta T.P$                               | -0.0004     | 0.0002          | -2.01 | 0.049  |
| Constant                                   | -0.0003     | 0.0554          | -0.01 | 0.996  |

---

1. $R^2$ value for Model 3 is 0.464
2. Robust standard error
3. T-statistic value – critical value 3.36 given 8 degrees of freedom in the Model at the 1% level.
4. Probability of statistical significant value – critical value 0.05 for the 95% confidence level.
5. $R^2$ value for Model 4 is 0.438
6. T-statistic value – critical value 4.60 given 4 degrees of freedom in the Model at the 1% level.

The $R^2$ value shows that the independent variables explained 46.4% and 43.8% of the influences that affect annual wheat yields in Models 3 and 4 respectively. This suggests weather has an influence on yields and so each variable will now be analysed individually.
Area
The p-value in both Models is less than the 0.05 critical value and so we can reject the null hypothesis at the 95% confidence level. Area is statistically significant and may be determining changes in yields. The coefficients indicate that a 1% increase in land cultivated causes a 0.8% increase in yields in both Models (when rounded to one decimal place), suggesting that marginal land productivity is growing. However, this assumes that there is access to new cultivatable land, which is not necessarily the case (GrainGrowers Ltd, 2011). Land availability is in decline across many parts of Australia, including Victoria.

Temperature
In Model 3, neither temperature term is statistically significant. However, the temperature change variable shows that a 1°C increase in temperature would decrease yields by 50%. If this pattern were linear, this would mean that a 2°C temperature increase (as predicted in many Climate Change Models) would eradicate wheat production from Victoria.

In Model 4, the squared temperature term is statistically significant and so indicates non-linear responses in yield to changes in temperature. This is to be expected as large increases (heatwaves) or decreases (cold periods) in temperature negatively impact the growth of crops. Figure 1 illustrates this non-linear relationship. Predicted values of the change in ln(Yield) for given temperature changes, along with the 95% confidence interval, are presented. A concave relationship exists, in line with the aforementioned expectations.

Figure 6: Effects of $\Delta T^2$ on $\Delta \ln Y$
The evidence from both Models highlights the threat that the climate poses to wheat production. Temperature increases clearly have a negative effect on yields in this study.

Precipitation

Again, neither precipitation variable is statistically significant in Model 3, but the variables offer an explanatory insight into the effect on yields. A 100mm increase in precipitation would cause a fall of 0.8% in yields. This mirrors the literature as floods cause crop nutrition loss. However, it was expected that precipitation changes would give non-linear results as both significant increases and decreases have negative effects.

The non-linear response in yields is illustrated in Figure 2 from results in Model 4. However, $P^2$ is not statistically significant in either specification in Table 4, but was statistically significant in Model specification 2 (see Appendix 2). Therefore, it may be inferred that large amounts of precipitation (floods) and small amounts (droughts) have a negative impact on yields. As precipitation is calculated using the growing season rainfall only, Figure 2 does not include periods of rainfall outside of this. This means that a flood or drought outside of the growing season is not considered. Whilst rainfall in November through to March (Australian summertime) does not directly affect growth of wheat, extreme weather events can affect, for example, soil nutrition which may have an impact on crops in the following growing season (van Rees et al., 2011).

Figure 7: Effects of $\Delta P^2$ on $\Delta \ln Y$
Temperature and Precipitation

Interaction terms were used to determine the relationship between variables. The temperature and precipitation interaction term in Model 4 is significant at the 5% level, suggesting that the effect of temperature is dependent on precipitation, and precipitation on temperature. Whilst these variables are not statistically significant in Model 3, they do show a stronger relationship. This reiterates the argument that both temperature and precipitation are important to the success of yields.

4.4.2 ET-SPI Model regression

Appendix 3 contains the regression results for a variety of Model specifications (1-11). This analysis will focus on the results presented in Table 6 (Model 7 and Model 9). As before, these have the best fit and greatest number of statistically significant variables respectively.

<table>
<thead>
<tr>
<th>Table 6: ET-SPI Model Regression Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ET-SPI Model Regression Results</strong></td>
</tr>
<tr>
<td><strong>Model 7 – dependent variable ΔlnY</strong></td>
</tr>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>ΔlnA</td>
</tr>
<tr>
<td>ΔETo</td>
</tr>
<tr>
<td>ΔETo$^2$</td>
</tr>
<tr>
<td>ΔSPI</td>
</tr>
<tr>
<td>ΔEToSPI</td>
</tr>
<tr>
<td>ΔETo$^2$SPI</td>
</tr>
<tr>
<td>Drought</td>
</tr>
<tr>
<td>Flood</td>
</tr>
<tr>
<td>ΔEToDrought</td>
</tr>
<tr>
<td>ΔEToFlood</td>
</tr>
<tr>
<td>Constant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Model 9 – dependent variable ΔlnYS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>ΔlnA</td>
</tr>
<tr>
<td>ΔEToETo</td>
</tr>
<tr>
<td>ΔSPISPI</td>
</tr>
<tr>
<td>ΔEToSPI</td>
</tr>
<tr>
<td>Drought</td>
</tr>
</tbody>
</table>
The R² values show that the independent variables explained 56% and 53.8% of the influences affecting annual wheat yields in Models 7 and 9, respectively. Each variable will now be analysed further.

**Area**

As for the T-P regression, area is statistically significant in both Models 7 and 9. However, these Models have coefficients indicating that a 1% increase in land cultivated causes a 1.1% increase in yields as opposed to 0.8% in the T-P Model. Again, land availability is questionable.

**ETo**

In none of the specified Models is ETo statistically significant, unless combined with an interaction term. This suggests that the evapotranspiration rate is not having an impact on yields. However, this contradicts both the literature and the T-P Model. There are two potential causes of this. The ETo equation was specified using solar radiation but a lack of data meant this term was transformed into an average (Equation 2). This could be affecting the ETo values that are generated and hence result in an insignificant coefficient being produced. Alternatively, the Hargraeves equation may be inappropriate for this study as compared to Blanc’s (2012) work. Further research into alternative ETo equations would be necessary, but is beyond the scope of this study.

**SPI, Drought and Flood**

In the results for Model 9, the SPI² term shows, again, a non-linear response in yields, and is statistically significant demonstrating the importance of precipitation. As the SPI calculates the deviation from the mean, it highlights the importance of fluctuations in precipitation during the growing season. Drought is also statistically significant highlighting how a lack of crop available moisture causes a reduction in yields. The coefficient is larger than for floods, supporting that Victoria is more affected by a lack of rainfall than too much.
Model 7, however, contradicts itself. The SPI coefficient suggests that extremely wet periods have a negative impact on yields whilst drier spells would not. Contrastingly, the drought dummy (which shows extremely dry SPI events) is statistically significant and gives a reduction in yields. The flood dummy is not statistically significant, but also shows an increase in yields during extremely wet events. This contradiction suggests that droughts have a greater influence on yields than floods in terms of extreme weather events, but that when assessing fluctuations around the mean, increases above the mean cause larger effects. This is not the result that was to be expected. However, it may be explained due to the fact that there were quite simply more droughts in 1953-2011 than floods, and that, for years outside of these extreme events, Victoria experienced greater intensity of rainfall making minor floods more common and troublesome. This is supported by the literature and so demonstrates the complexity of rainfall and the resulting impact on yields.

Despite soil nutrition being excluded from this study, the significance of droughts suggests that this could be an area for further research to increase the strength of the Model.

**ETo and SPI**

The interaction of ETo and SPI has a large coefficient in Model 7 and is statistically significant in Model 9. This shows that temperature and rainfall depend on one another for their effects, which was to be expected given the results from the T-P Model.

### 4.4.3 Summary of Results

The regression results from the T-P and ET-SPI Models partly support the hypothesis for this study, arguing that large increases or decreases in temperature and precipitation have a negative effect on wheat yields in Victoria. Strong coefficients and statistical significance for squared terms in both Models highlight the non-linear relationship between weather and yields. However, some variables that were not statistically significant raise queries about the strength of the calculations for each variable, notably ETo.

### 4.5 Limitations and Further Research

One of the main limitations of this study is the exclusion of variables – influenced by weather patterns – that have an impact on wheat yields. Different fertilisers are selected based on rainfall and temperature, for example (GrainGrowers Ltd, 2011). However, inclusion of these variables would be extremely difficult on this scale due to the lack of data available.
Additionally, crop rotation is not considered which could greatly affect yields. Not all farms practice crop rotations and again data availability is limited (GrainGrowers Ltd, 2011). However, if major farms do not produce wheat for one growing season, this could affect the Model to a currently unknown extent.

Due to Model limitations, SPI has to be calculated using a 6-month average, which means one month of the growing season is lost in the calculation (WMO, 2012). This is not optimal when the focus of this study is weather alone. The alternative – to run monthly SPIs and track averages from this – would require obtaining daily data for each month which is again outside the scope of this study and would raise data limitation issues due to the length of the time period focused on (BoM, 2015).

Further research including variables that are determined by weather – soil quality, fertiliser choice, and crop management techniques to name but a few – could strengthen the Model and demonstrate the sheer scale of the wheat growing industry’s reliance on weather patterns. Previous studies have built Models predicting future yields by incorporating Climate Change scenarios (DAFF, 2006; GrainGrowers Ltd, 2011; Linehan et al., 2012; Wang et al., 1992). These studies do not, however, incorporate more complex measures such as ETo and SPI. Hence building on this study by using Climate Change scenarios could be a direction for future research.

5. Conclusion

This study has sought to explain the impacts that weather has had on wheat yields in Victoria, Australia. The importance of the wheat industry was highlighted in the second Section. The success of the industry has a profound impact on the economy. Periods of drought in Victoria are increasingly prevalent, the impacts of which were to be a focal point in this study.

Section 3 highlights that substantial research has been conducted to assess how temperature changes affect Australia. Direct and indirect impacts are a cause for concern. This study focused on the indirect effect of increased water demand. The literature provided the fundamentals of understanding more complex temperature measures in the Australian context (Wang et al., 1992), including evapotranspiration. By adapting Blanc’s (2012) work, and combining it with Australian research, this study managed to apply new measures that expanded on the current body of literature.

The same was achieved for rainfall. Across the country, Ludwig and Asseng (2006) found that precipitation had non-linear effects. The limited studies on Victoria concluded that rainfall is
decreasing in the growing season, which has impacted wheat crops (GrainGrowers Ltd, 2011). Of most concern, this decrease regularly coincides with years of higher average temperatures (Anwar et al., 2007). Again, this study was able to expand on the literature by using a more complex measure (SPI) that has the ability to isolate extreme weather events from the averages.

Consequently, Section 3 allowed for the identification of opportunities in the current body of literature that were then incorporated into the empirical analysis. Section 4 explained the background to each variable considered. This was not without its limitations. Notably, the decisions to use a 6-month SPI and to develop an adjusted ETo equation created weaknesses in the study.

The Model specifications and subsequent regressions were based on Blanc’s (2012) study on sub-Saharan Africa. This was adapted to fit the Australian context and Victorian dataset in this study. Consequently, a number of specifications were developed for each Model to consider the different options. The strength of these varied, but the Model of best fit and the Model with the most statistically significant variables were chosen for the analysis. All four specifications considered showed that yields increase as area increases. This does not consider land availability decline. Hence a solution to falling yields is not necessarily to cultivate more land exponentially.

Overall, the results for temperature corroborated the results of the current literature – increases in temperature cause declines in crop yield. However, the relationship was proven to be non-linear in the T-P Model – a fall in temperatures can reduce yields. The background data suggests that this is not a cause for concern in the long-term given the continued increase in temperature on average over the past 100 years. Additionally, the ETo was not statistically significant in the ET-SPI Model, highlighting a potential weakness in the adjusted equation used.

Similarly, precipitation had non-linear effects in the T-P Model. Combining this result with the SPI from the ET-SPI Model, it can be concluded that drought has a greater impact on yields than periods of increased rainfall. Interaction of the temperature-precipitation and ETo-SPI variables demonstrate the complexity of the effects. Again, this reflects the views widely held in the current body of literature whilst presenting a myriad of opportunities for future study.

Key limitations of this study can be addressed through further research. Alternative ET equations are available and could be tested to increase the statistical significance of the variable and SPI may be calculated on a monthly basis to account for the entire GSR. However, this would require extensive
data sourcing beyond the scope of this study. Lack of data availability also prevented the use of variables such as soil nutrition and crop rotation. Including these in the Model, and using Climate Change predictions to estimate the future, is a direction that this study could take. The opportunity for expansion is both vast and relevant to the current body of literature.

Overall, this research achieved its objective. Weather changes in the study period have negatively impacted wheat yields in the state of Victoria. The results highlight the increased risk that the future holds because of Climate Change.
Bibliography


Appendices

Appendix 1: Wheat Production Data for Victoria, 1953-2011

**Figure 8: Area under Wheat Production (hectares)**

Source(s): Author's calculations; Data Source: ABS (2013)

**Figure 9: Output of Wheat Production (tonnes)**

Source(s): Author’s calculations; Data Source: ABS (2013)
### Appendix 2: T-P Model regression results of all specified Models

**Table 7: T-P Model regression results of all specified Models**

<table>
<thead>
<tr>
<th></th>
<th>Model specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta \ln A )</td>
<td></td>
<td>0.848</td>
<td>0.782</td>
<td>0.776</td>
<td>0.838</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.391)</td>
<td>(0.386)</td>
<td>(0.385)</td>
<td>(0.393)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.035]</td>
<td>[0.048]</td>
<td>[0.049]</td>
<td>[0.038]</td>
</tr>
<tr>
<td>( \Delta T )</td>
<td></td>
<td>0.590</td>
<td>-6.061</td>
<td>-5.031</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.292)</td>
<td>(5.256)</td>
<td>(19.024)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.048]</td>
<td>[0.254]</td>
<td>[0.793]</td>
<td></td>
</tr>
<tr>
<td>( \Delta T^2 )</td>
<td></td>
<td>0.204</td>
<td>0.178</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.184)</td>
<td>(0.694)</td>
<td>(0.005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.275]</td>
<td>[0.798]</td>
<td>[0.009]</td>
<td></td>
</tr>
<tr>
<td>( \Delta P )</td>
<td></td>
<td>0.012</td>
<td>-0.059</td>
<td>-0.008</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.009)</td>
<td>(0.029)</td>
<td>(0.376)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.204]</td>
<td>[0.045]</td>
<td>[0.982]</td>
<td></td>
</tr>
<tr>
<td>( \Delta P^2 )</td>
<td></td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.001]</td>
<td>[0.829]</td>
<td>[0.069]</td>
<td></td>
</tr>
<tr>
<td>( \Delta TP )</td>
<td></td>
<td>-0.001</td>
<td>0.003</td>
<td>-0.002</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.001)</td>
<td>(0.002)</td>
<td>(0.052)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.180]</td>
<td>[0.097]</td>
<td>[0.969]</td>
<td>[0.049]</td>
</tr>
<tr>
<td>( \Delta T^2 P )</td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.954]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta TP^2 )</td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.718]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td></td>
<td>0.000</td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.056)</td>
<td>(0.055)</td>
<td>(0.057)</td>
<td>(0.055)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.996]</td>
<td>[1.000]</td>
<td>[0.997]</td>
<td>[0.996]</td>
</tr>
<tr>
<td><strong>Model R^2 value</strong></td>
<td></td>
<td>0.419</td>
<td>0.460</td>
<td>0.464</td>
<td>0.438</td>
</tr>
</tbody>
</table>

*Note(s): ( ) robust standard errors; [ ] probability values*
**Appendix 3: ET-SPI Model regression results of all specified Models**

**Table 8: ET-SPI Model regression results of all specified Models**

<table>
<thead>
<tr>
<th>Model specification</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔlnA</td>
<td>0.932</td>
<td>0.935</td>
<td>0.9927</td>
<td>0.891</td>
</tr>
<tr>
<td></td>
<td>(0.389)</td>
<td>(0.408)</td>
<td>(0.404)</td>
<td>(0.371)</td>
</tr>
<tr>
<td></td>
<td>[0.020]</td>
<td>[0.026]</td>
<td>[0.026]</td>
<td>[0.020]</td>
</tr>
<tr>
<td>ΔETo</td>
<td>0.823</td>
<td>-0.814</td>
<td>-6.559</td>
<td>0.664</td>
</tr>
<tr>
<td></td>
<td>(0.932)</td>
<td>(39.196)</td>
<td>(38.348)</td>
<td>(1.072)</td>
</tr>
<tr>
<td></td>
<td>[0.381]</td>
<td>[0.984]</td>
<td>[0.865]</td>
<td>[0.538]</td>
</tr>
<tr>
<td>ΔETo²</td>
<td>0.479</td>
<td>2.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(11.394)</td>
<td>(11.139)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.967]</td>
<td>[0.856]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔSPI</td>
<td>0.853</td>
<td>0.786</td>
<td>-8.665</td>
<td>1.073</td>
</tr>
<tr>
<td></td>
<td>(0.956)</td>
<td>(1.933)</td>
<td>(13.328)</td>
<td>(1.311)</td>
</tr>
<tr>
<td></td>
<td>[0.377]</td>
<td>[0.686]</td>
<td>[0.519]</td>
<td>[0.417]</td>
</tr>
<tr>
<td>ΔSPI²</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔEToSPI</td>
<td>-0.591</td>
<td>-0.005</td>
<td>10.259</td>
<td>-0.703</td>
</tr>
<tr>
<td></td>
<td>(0.568)</td>
<td>(0.056)</td>
<td>(15.596)</td>
<td>(0.795)</td>
</tr>
<tr>
<td></td>
<td>[0.303]</td>
<td>[0.931]</td>
<td>[0.514]</td>
<td>[0.380]</td>
</tr>
<tr>
<td>ΔETo²SPI</td>
<td></td>
<td></td>
<td>-3.087</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(4.583)</td>
<td>[0.504]</td>
</tr>
</tbody>
</table>

| Drought        | 0.026|      |      |      |
|                | (0.166)|      | [0.877]| |
| Flood          |      |      | -0.114|      |
|                |      |      | (0.173)| [0.512]| |
| ΔEToDrought    |      |      |      |      |

*Table continued on next page*
<table>
<thead>
<tr>
<th>ΔEToFlood</th>
<th>[ \bar{0}.005 ]</th>
<th>[ \bar{0}.005 ]</th>
<th>[ \bar{0}.006 ]</th>
<th>[ \bar{0}.004 ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>(0.056)</td>
<td>(0.056)</td>
<td>(0.057)</td>
<td>(0.057)</td>
</tr>
<tr>
<td></td>
<td>[0.931]</td>
<td>[0.931]</td>
<td>[0.917]</td>
<td>[0.945]</td>
</tr>
<tr>
<td>Model R^2 value</td>
<td>0.426</td>
<td>0.426</td>
<td>0.431</td>
<td>0.430</td>
</tr>
</tbody>
</table>

Note(s): ( ) robust standard errors; [ ] probability values

*Appendix 3 continued on next page*
<table>
<thead>
<tr>
<th>Model specification</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔlnA</td>
<td>1.127</td>
<td>0.895</td>
<td>1.094</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>(0.391)</td>
<td>(0.377)</td>
<td>(0.371)</td>
<td>(0.404)</td>
</tr>
<tr>
<td></td>
<td>[0.006]</td>
<td>[0.022]</td>
<td>[0.005]</td>
<td>[0.038]</td>
</tr>
<tr>
<td>ΔETo</td>
<td>0.993</td>
<td>-3.221</td>
<td>3.732</td>
<td>-21.405</td>
</tr>
<tr>
<td></td>
<td>(0.930)</td>
<td>(48.308)</td>
<td>(39.704)</td>
<td>(60.272)</td>
</tr>
<tr>
<td></td>
<td>[0.291]</td>
<td>[0.947]</td>
<td>[0.926]</td>
<td>[0.724]</td>
</tr>
<tr>
<td>ΔETo^2</td>
<td>1.049</td>
<td>-0.726</td>
<td>6.155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(14.153)</td>
<td>(11.636)</td>
<td>(17.558)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.941]</td>
<td>[0.950]</td>
<td>[0.727]</td>
<td></td>
</tr>
<tr>
<td>ΔSPI</td>
<td>1.973</td>
<td>-6.244</td>
<td>-0.402</td>
<td>18.171</td>
</tr>
<tr>
<td></td>
<td>(0.916)</td>
<td>(13.999)</td>
<td>(11.461)</td>
<td>(30.790)</td>
</tr>
<tr>
<td></td>
<td>[0.036]</td>
<td>[0.658]</td>
<td>[0.972]</td>
<td>[0.558]</td>
</tr>
<tr>
<td>ΔSPI^2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔEToSPI</td>
<td>7.625</td>
<td>1.174</td>
<td>-21.585</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.640]</td>
<td>[0.930]</td>
<td>[0.554]</td>
<td></td>
</tr>
<tr>
<td>ΔETo^2SPI</td>
<td>-2.368</td>
<td>-0.579</td>
<td>6.337</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.733)</td>
<td>(3.843)</td>
<td>(10.684)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.619]</td>
<td>[0.881]</td>
<td>[0.556]</td>
<td></td>
</tr>
<tr>
<td>Drought</td>
<td>-0.582</td>
<td>-0.893</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.216)</td>
<td>(0.251)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.009]</td>
<td>[0.001]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flood</td>
<td>0.028</td>
<td>0.075</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.120)</td>
<td>(0.193)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.814]</td>
<td>[0.698]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔDrought</td>
<td>0.026</td>
<td></td>
<td>-8.376</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.197)</td>
<td></td>
<td>(5.836)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.894]</td>
<td></td>
<td>[0.158]</td>
<td></td>
</tr>
<tr>
<td>ΔFlood</td>
<td>-0.080</td>
<td>-6.654</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.191)</td>
<td></td>
<td>(10.724)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.675]</td>
<td></td>
<td>[0.538]</td>
<td></td>
</tr>
<tr>
<td>ΔEToDrought</td>
<td></td>
<td>0.276</td>
<td>4.771</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.124)</td>
<td>(3.389)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.031]</td>
<td>[0.166]</td>
<td></td>
</tr>
<tr>
<td>ΔEToFlood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.030</td>
<td>4.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.144)</td>
<td>(6.503)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.834]</td>
<td>[0.540]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.077</td>
<td>-0.005</td>
<td>0.120</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
<td>(0.058)</td>
<td>(0.076)</td>
<td>(0.059)</td>
</tr>
<tr>
<td></td>
<td>[0.266]</td>
<td>[0.933]</td>
<td>[0.120]</td>
<td>[0.901]</td>
</tr>
<tr>
<td>Model R² value</td>
<td>0.510</td>
<td>0.433</td>
<td>0.560</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note(s): ( ) robust standard errors; [ ] probability values

*Appendix 3 continued on next page*
<table>
<thead>
<tr>
<th>Variable</th>
<th>(9)</th>
<th>(10)</th>
<th>(11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta \ln A )</td>
<td>1.108</td>
<td>1.102</td>
<td>1.129</td>
</tr>
<tr>
<td></td>
<td>(0.384)</td>
<td>(0.359)</td>
<td>(0.391)</td>
</tr>
<tr>
<td></td>
<td>[0.006]</td>
<td>[0.003]</td>
<td>[0.006]</td>
</tr>
<tr>
<td>( \Delta E_{\text{To}} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta E_{\text{To}}^2 )</td>
<td>0.456</td>
<td>0.388</td>
<td>0.285</td>
</tr>
<tr>
<td></td>
<td>(0.279)</td>
<td>(0.287)</td>
<td>(0.272)</td>
</tr>
<tr>
<td></td>
<td>[0.109]</td>
<td>[0.182]</td>
<td>[0.300]</td>
</tr>
<tr>
<td>( \Delta SPI )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta SPI^2 )</td>
<td>1.175</td>
<td>1.932</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.928)</td>
<td>(0.921)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.211]</td>
<td>[0.041]</td>
<td></td>
</tr>
<tr>
<td>( \Delta E_{\text{To}} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta E_{\text{To}} )</td>
<td>-0.090</td>
<td>-0.738</td>
<td>-1.251</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.564)</td>
<td>(0.552)</td>
</tr>
<tr>
<td></td>
<td>[0.049]</td>
<td>[0.197]</td>
<td>[0.028]</td>
</tr>
<tr>
<td>( \Delta E_{\text{To}}^2 )</td>
<td>-0.725</td>
<td>-0.897</td>
<td>-0.581</td>
</tr>
<tr>
<td>( Drought )</td>
<td>(0.214)</td>
<td>(0.235)</td>
<td>(0.215)</td>
</tr>
<tr>
<td></td>
<td>[0.001]</td>
<td>[0.000]</td>
<td>[0.009]</td>
</tr>
<tr>
<td>( Flood )</td>
<td>-0.027</td>
<td>0.071</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>(0.115)</td>
<td>(0.193)</td>
<td>(0.120)</td>
</tr>
<tr>
<td></td>
<td>[0.816]</td>
<td>[0.716]</td>
<td>[0.818]</td>
</tr>
</tbody>
</table>

Table continued on next page
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta E_{ToFlood} )</td>
<td>-0.032</td>
<td>(0.136)</td>
<td>[0.814]</td>
</tr>
<tr>
<td>Constant</td>
<td>0.111</td>
<td>0.122</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.071)</td>
<td>(0.069)</td>
</tr>
<tr>
<td></td>
<td>[0.104]</td>
<td>[0.091]</td>
<td>[0.267]</td>
</tr>
<tr>
<td>Model R^2 value</td>
<td>0.538</td>
<td>0.559</td>
<td>0.510</td>
</tr>
</tbody>
</table>

Note(s): ( ) robust standard errors; [ ] probability values
The impact of quantitative easing on financial markets in the United Kingdom

James Barrie

Following the global financial crisis, the Bank of England was forced to take dramatic measures in an attempt to improve the UK economy. In March 2009, the Bank of England launched an unprecedented, large-scale asset-purchasing programme called quantitative easing. By October 2012, the Bank had bought £375bn worth of assets, primarily government bonds. This paper attempts a provisional evaluation of the impact of quantitative easing on financial markets in the UK and, to a lesser extent, the economy more generally. At this point in time, it is very difficult to assess these impacts by conventional means. As a result, prior studies have typically used event study analysis in attempts to quantify the policy’s impact. After reviewing the event study literature, the findings from the paper’s own qualitative survey of financial market experts are presented. We are able to draw two main conclusions. Firstly, the policy has had a significant effect on financial markets in much the way that was both intended and expected. In so doing, it is believed that quantitative easing will also have played an important role in stabilising the economy following the recession. This study points to a negative effect of around 50 basis points on gilt yields. However, it remains difficult to estimate the exact size of the effect that the policy has had. Secondly, the evidence suggests that the policy’s main influence came through a portfolio rebalancing channel. As always, there is more research that can, and will, be done and it’s hoped that the survey-based approach adopted by this paper can be extended in the future.
1. Introduction

The global financial crisis that started in 2007 was led by the sub-prime mortgage crisis in the US, but quickly became more generalised. In September of that year, the UK mortgage lender Northern Rock became the first British casualty, requiring intervention in the form of emergency funding from the Bank of England. A year or so later, events took a significant turn for the worse as one of the world’s largest banks, the US investment bank Lehman Brothers, collapsed. Confidence in the global financial system and the global economy fell sharply and markets became dysfunctional, limiting liquidity and the availability of funding. A month later, in October 2008, the UK Government was forced to bail out and take a stake in three of the larger British lenders – RBS, Lloyds TSB and HBOS (BBC, 2008).

The recession that followed the financial crisis was the most severe and synchronised economic downturn in modern history. In the first quarter of 2009, UK manufacturing output fell at an annualised rate of close to 20% and unemployment rose at an unprecedented rate (Dale, 2010). In March, the FTSE 100 stock market index was almost 50% lower than it had been only eighteen months earlier (see Chart 4). In the second quarter, the UK economy was almost 6% smaller than it had been a year earlier - the sharpest fall ever recorded (see Appendix 1). The pre-recession level of activity has only been regained within the past year.

The extraordinary combination of financial crisis and recession presented a major challenge to governments and central banks around the world. It’s no surprise, therefore, that the policy response was both unprecedented and unconventional. In the UK, the Bank of England’s monetary policymaking body - the Monetary Policy Committee (MPC) - started to cut the bank rate, the key short-term interest rate at which commercial banks borrow from the central bank, in October 2008 (Inman, 2015). In March of the following year, the bank rate was cut for the sixth successive month to 0.5%, the lowest level in the three-hundred-year history of the Bank (Butt et al., 2012). This was seen as the effective floor for short rates and, hence, the effective limit to conventional monetary policy. However, it was also evident that these measures, and the expansionary fiscal policy represented by the increase in the budget deficit, might not prove sufficient (Dale, 2010). At the same time that the limit to conventional monetary policy was reached, the MPC embarked on an additional, unconventional form of monetary easing.

The term ‘quantitative easing’ was first applied to the Bank of Japan’s purchases of government securities from the banking sector, which had been aimed at fighting off deflation
in the early 2000s (Joyce et al., 2012). In more recent times, the Federal Reserve, the Bank of England, and more recently still, the European Central Bank have all undertaken their own quantitative easing programmes. In the UK, quantitative easing was launched in March 2009. The policy involves taking assets onto the central bank’s balance sheet and increasing the money supply as a result. The balance sheet and the money supply are the ‘quantities’ in ‘quantitative easing’. In particular, the Bank of England set out to buy medium and long-term government bonds or gilts from the non-bank private sector (Benford et al., 2009). Through several possible transmission channels, or ‘channels of influence’, the intention was that the expansion of the broad money supply should boost spending and income, helping to stabilise growth and inflation. The MPC initially set out to buy £75bn of gilts. However, the programme quickly grew in size and in 2012 the total purchases of government bonds rose to £375bn (Inman, 2015). The MPC has subsequently maintained the stock at this level by reinvesting cash flows from maturing assets (Bank of England, 2014b).

This paper attempts a provisional evaluation of the impact of quantitative easing in the UK. We will review the literature on the topic and present our own survey in order to identify the important channels of influence and evaluate the policy’s impact on financial markets and, to a lesser extent, the economy more generally. On the latter, it is hard to accurately estimate the policy’s impact on the economy due to the difficulty in accounting for the coincident effects of other policy measures, not least conventional monetary and fiscal policy, as well as the impact of other influences (Joyce et al., 2011). For the most part, then, this paper looks at the impact on financial markets, including the gilt market. This provides an indication of the initial impact and we would suggest that the effectiveness of the policy in addressing the ultimate targets – spending and inflation – is likely to depend at least in part on its impact on the immediate targets.

The structure of the remainder of the paper is as follows. Section 2 describes the main features of the UK programme and its implementation. Section 3 sets out the main channels through which the policy might be expected to have an impact. Sections 4 and 5 then look at the impact of quantitative easing on financial markets; in Section 4, we present and discuss some of the relevant time-series data, and in Section 5 we review the recent event study analysis. Such studies analyse the market reaction immediately after each announcement or each round of quantitative easing in an attempt to hold other things equal. Our own qualitative survey of commentators, financial market practitioners and former policy-makers in Section 7 provides a different method of evaluating quantitative easing. In its own way, it’s
also an attempt to hold other things equal. In our view, the survey results complement the event study analysis and strengthen the conclusions of the paper presented in the final section.

2. Implementation

Since January 2009, in the UK, money supply injections via the purchasing of assets have been conducted by a wholly owned subsidiary of the Bank of England, called the Asset Purchase Facility (APF). Full indemnity is provided by HM Treasury to protect the Bank of England against any losses, but the APF is independent of the government (Bean, 2009). Initially, at the start of 2009, the APF was authorised to buy up to £50bn high quality private sector assets—corporate bonds and commercial paper (Joyce, Tong and Woods, 2011). These purchases were aimed at reducing the illiquidity of credit markets in order to increase the available credit for firms. The APF began buying assets in February, however, market functioning was restored reasonably quickly and the quantities of assets purchased never came close to the £50bn ceiling (Fawley and Neely, 2013).

As the economic situation worsened, in March 2009, the MPC announced another programme of asset purchases to be undertaken by the APF that, crucially, was to be financed by central bank reserves. This was the start of quantitative easing (Joyce, Tong and Woods, 2011). Initially, it was proposed that £75bn worth of assets would be purchased over three months. The purchases consisted mainly of UK government bonds, otherwise known as ‘gilts’. Policy-makers argued that gilt purchases from the non-bank private sector would increase broad money growth without incurring credit risk. In order to target assets held by the non-bank private sector most effectively, gilts with maturity dates between 5 and 25 years were purchased. This was because banks often hold very short-dated gilts, whereas longer-dated gilts are typically held by private institutions matching long-dated liabilities (Joyce, Tong and Woods, 2011).

The assets were purchased through reverse auctions. The Bank of England periodically announced schedules of auctions, usually held two or three times a week, along with the maturity range and quantity of gilts to be purchased (Treasury, 2011). In a reverse auction, bonds are purchased from the lowest bidders rather than sold to the highest bidders. So, in the quantitative easing auctions sellers submitted prices at which they were prepared to sell specific quantities of gilts. The APF then accepted the lowest offers until it had bought the total amount stated for that auction (Benford et al., 2009). The sellers participating in the
auction were banks and securities dealers who were permitted to submit bids on behalf of their clients (Benford et al., 2009).

The implementation of quantitative easing in the UK can be divided into two distinct periods. Following the launch of the policy in March 2009, the size of the programme was increased in May, and again in August, and then finally to £200bn in November. It remained at this level for almost two years. In October 2011, the programme was expanded as the medium-term outlook for growth and inflation in the UK deteriorated, partly as a result of growing concerns over the euro area economy (Inman, 2015). The programme was subsequently expanded again in February 2012, and then again, finally, in July of that year to £375bn. The stock of assets has been maintained at this level ever since. Table 1 lists the important announcements in more detail (Joyce, Tong and Woods, 2012; Fawley and Neely, 2013).
### Table 1: Quantitative Easing in the UK: Key Announcements

<table>
<thead>
<tr>
<th>Date</th>
<th>Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>February 11th 2009</td>
<td>The Bank of England’s February Inflation Report gave a strong indication that a quantitative easing programme was likely.</td>
</tr>
<tr>
<td>March 5th 2009</td>
<td>The MPC cut bank rate to 0.5%. It also announced it would purchase an initial £75bn of assets financed by central bank reserves over the next 3 months. Gilt purchases were to be restricted to bonds with a maturity of between 5 and 25 years.</td>
</tr>
<tr>
<td>May 7th 2009</td>
<td>The MPC announced that the size of the asset purchase programme would be increased to £125bn.</td>
</tr>
<tr>
<td>August 6th 2009</td>
<td>The programme was extended again by the MPC to £175bn. The decision was made to target all gilts with a maturity of over 3 years in order to not further disrupt the 5 to 25 year gilt market.</td>
</tr>
<tr>
<td>November 5th 2009</td>
<td>The MPC announced that the quantitative easing programme would be increased to £200bn.</td>
</tr>
<tr>
<td>February 4th 2010</td>
<td>The MPC announced that the stock of asset purchases would be maintained at £200bn.</td>
</tr>
<tr>
<td>October 6th 2011</td>
<td>The MPC announced that the quantitative easing programme would be expanded to £275bn over the following 4 months.</td>
</tr>
<tr>
<td>February 9th 2012</td>
<td>A further £50bn of quantitative easing was announced by the MPC, taking the total stock to £325bn.</td>
</tr>
<tr>
<td>July 5th 2012</td>
<td>The MPC announced that the quantitative easing programme would increase to £375bn.</td>
</tr>
</tbody>
</table>

Source(s): Joyce, Tong and Woods (2011); and Fawley and Neely (2013)

### 3. Transmission Mechanisms

#### 3.1 Impact on the money supply

The fundamental aim of the policy is to boost spending by, among other things, increasing the money supply. In order to directly increase the broad money supply, as measured by the monetary aggregate M4, asset purchases must be associated with an increase in the bank deposits held by the non-bank private sector (Butt et al., 2012). Asset purchases made from the banking sector, or that lead to non-banks repaying liabilities, won’t necessarily increase the money supply, though they might have helpful impacts of a different sort. Any increase in
broad money may have an impact on the economy through a number of channels. On implementing quantitative easing, policy-makers were uncertain about the strength and the timing of the different channels due to the unprecedented nature of the policy (Benford et al., 2009). Some have argued that describing the process in any great detail is unnecessary and that an increase in broad money should, by definition, increase economic activity or prices (Benford et al., 2009). This is the quantity theory of money. In practice, while M4 growth has been positively correlated with nominal growth in the UK, the relationship has not been particularly strong and subject to so-called long and variable lags (Benford et al., 2009). As the programme has continued, discussion and research has tended to focus on three key transmission channels as identified and discussed below.

3.2 Portfolio rebalancing

Under certain assumptions, large-scale asset purchases should raise the price and reduce the yield of the assets in question. On that basis, a programme such as that in the UK that focuses on government bonds can be expected to reduce gilt yields. With short rates at the effective floor, the policy can be thought of as acting on long-term rates. In addition, the initial effect on gilt yields is likely to initiate a broader chain of transactions and revaluations that raise prices and reduce yields on a wider range of assets. In part, this is because some of those who sell gilts are likely to view the money obtained as a poor substitute. To rebalance their portfolios, they may use their excess money balances to purchase other assets, such as equities or corporate bonds (Benford et al., 2009). The money is then transferred to the sellers of those assets who may similarly need to buy other assets to rebalance their own portfolios, and on it goes. The rise in other asset prices rise, and corresponding falls in yields, may lead to an increase in the issuance of new assets and finance for new spending. The rise in asset prices also increases the wealth of existing asset holders, which may boost their spending, though some may choose to save more in response to lower rates (Joyce et al., 2011). This is the channels through which the MPC believed quantitative easing would have the most impact (Benford et al., 2009).

3.3 Impact on the exchange rate

On the same assumption that investors view money as an imperfect substitute for non-money assets, and are likely to purchase other assets as a result of quantitative easing, it’s possible that some of the new purchases may include foreign assets, including foreign government bonds (Joyce et al., 2011). Furthermore, as a result of the decline in yields of a wider range of assets from the portfolio balancing effect, investors may exchange UK assets for foreign ones.
in search of higher returns. Therefore, the policy could put downward pressure on the sterling exchange rate, making UK exports more competitive and thus increasing demand (Joyce et al., 2011). Another implication of this point is that some of the impact – possibly a relatively small part – of the UK programme may have leaked overseas to other economies, and vice versa in the case of the US and Japanese programmes (Butt et al., 2012).

3.4 Impact on short rate expectations

Large-scale asset purchases are also likely to affect market participants’ expectations about future monetary policy. A programme of quantitative easing suggests – and can therefore be used to signal – that short rates are likely to remain low for an extended period of time (Joyce et al., 2011). If all else is equal, this should add to the effectiveness of conventional monetary policy. For some commentators, particularly those in the US, this ‘signalling’ effect represents a key aspect of the transmission mechanism (Joyce et al., 2012).

From an analytical point of view, there is an observational equivalence here in that the portfolio rebalancing effect and any impact on short rate expectations are both likely to be associated with lower gilt yields. As we shall see, additional information from other markets – including short rate derivatives – may allow us to separate the two to some extent, but they are clearly closely related in nature and effect.

4. Time-series data on the impact of quantitative easing

As noted above, there are a number of challenges in evaluating the impact of quantitative easing. The episode is unique in a UK context as is the simultaneous event of having short rates at unprecedentedly low levels. The programme is also still very recent, if not contemporary, which means the full impact may not have been seen yet. From an econometric point of view, this means that measuring its impact is very difficult; it’s hard to identify all of the other relevant variables and there are very few policy-on and subsequent policy-off observations on which to base estimates. Furthermore, if the stock of asset purchases rather than the flow is the key parameter, then the policy is still in place. In addition, given that we normally assume that monetary policy works with a lag, it’s unlikely that we have seen the full impact of the programme. In the future, it may be possible to undertake robust time-series analysis of the UK data, and also to make cross-country comparisons of the experience of the UK with that of the US and Japan, where a similar policy took place at a similar time, and the euro area, where it has only been introduced more recently. For now,
however, initial attempts of evaluation have tended to focus on the initial impact on financial markets and asset prices rather than the subsequent and ultimate impact on the economy.

In the sections that follow, we start by charting some key financial market and asset price indicators in order to form an impression of their behaviour during the period in which asset purchases were made. We go on to review the event study-type literature on the subject and, as a further form of evaluation, present the findings from our own expert survey into the attitudes and opinions of a sample of commentators and financial market practitioners.

The following charts plot ten years of daily data to help assess the impact of quantitative easing on the gilt market, the UK equity market and the exchange rate. These are the markets that one might expect to be most affected by the programme and, as outlined in Section 3, they are potentially key parts of the transmission mechanism. The shaded area on the charts covers the period between February 2009, when the programme was first announced, and the final round of asset purchases some three and a half years later at the end of Q3 2012.

Let us start with 10-year nominal gilt yields, illustrated in Chart 1. In very general terms, the chart shows that they are lower after the programme than before, by around 2% or 200 basis points. However, the movement within the shaded period is inconsistent and there was a sizeable, albeit temporary, rise in yields after the final round of asset purchases. The nominal gilt yield is made up of two components: the real yield, and an implied inflation component, which is a measure of the market’s inflation expectations. Chart 2 and Chart 3 use data from index-linked gilts to decompose the change in nominal yields into these two components. There are two issues here. First, the changing balance of supply and demand in the gilt market during the programme might be expected to affect real yields more than inflation expectations. Second, and in addition, a successful asset purchase programme might reasonably be expected to raise inflation expectations, particularly if, as was the case, the starting-point was one in which there was a perceived risk of deflation (Benford et al., 2009).
Chart 1: 10-year nominal gilt yield (%)

Chart 2: 10-year real gilt yield (%)

Source(s): this paper’s own illustration of Bank of England statistics, April 2015
The charts suggest that both issues may be relevant. The fall in real yields in the highlighted area is of roughly the same size as the fall in nominal yields at 200 basis points, but is more consistent and extends to the point at which real yields became – and still are – negative. It also appears that the programme may have helped stabilise inflation expectations. A sharp fall in implied inflation before the programme started was reversed when the asset purchases were announced. It’s clear that some of the volatility in nominal yields during the programme can be attributed to changes in implied inflation.

The trend in the UK equity market, illustrated in Chart 4, is similar to that observed in implied inflation. A fall in the equity market before the programme started was reversed when it got underway. The equity market has continued to rise more recently, reaching a high earlier this year at a level roughly double the low point seen just before quantitative easing started.

Turning to the exchange rate, Chart 5 shows the sterling-dollar cross rate, Chart 6 shows the euro-sterling cross rate, and Chart 7 shows the sterling effective exchange rate. Cross rates show the nominal exchange rate between two currencies, and the effective exchange rate measures the strength of the sterling relative to a basket of other currencies. The patterns on the main cross rates are rather different. Sterling weakened against both the dollar and the euro ahead of the programme. Since the quantitative easing started, Sterling remained closer
to the weaker end of its ten-year range against the dollar, but did return to the stronger end against the euro. As always, exchange rate movements reflect developments on both sides. The observed pattern is consistent with the expectation that the US is likely to tighten monetary policy before the UK, and that the euro area would require further monetary easing (Miller, 2015).

**Chart 4: FTSE 100 (‘000s)**

![FTSE 100 Chart](source)

Chart 5: *Sterling-dollar cross rate*

![Sterling-dollar Cross Rate Chart](source)
We must be cautious in drawing conclusions from this evidence. To state the obvious, observations made during the period in which asset purchases were made cannot necessarily be attributed to the asset purchases themselves. It’s possible that conventional monetary policy, in the form of the extended period of very low short-term rates, or, alternatively,
various other developments in the global economy, are responsible for what we see in the charts. On the other hand, though the relationship is not conclusive or quantified, the evidence is at least consistent with the programme having had an impact in the direction that might have been expected.

In addition, we also examined the largest ten daily moves in the same financial markets and asset prices across the ten years of data. The law of averages would suggest that a period lasting around three and a half years would contain three or four of the largest daily moves in each of these markets given the ten-year timeframe. However, if quantitative easing is determined to have had a significant effect then one may expect to observe more of the largest moves between February 2009 and October 2010. In fact, nine of the ten largest daily changes in gilt yields occurred during the period of asset purchases. In contrast, all but one of the largest moves in the FTSE 100 index occurred outside the programme and all of the largest moves in the effective sterling exchange rate took place before it started. This provides further evidence – again inconclusive – that quantitative easing may have had a significant impact on gilt yields, and, albeit one that was less pronounced, or possibly simply more gradual, in the case of other financial markets and asset prices.

So much is unusual about the period that many other factors and influences could be responsible for these observations. The usual econometric approach would be to reflect those factors and influences in a series of explanatory variables in a multivariate regression. As we explained earlier, however, this is not possible at present so we need to consider alternative methods of analysis. Event studies have proved to be the most popular analytical approach amongst the literature, including research undertaken by authorities such as the Bank of England who have sought to evaluate their own policies. An event study can be defined as a method of analysis that assesses the reactions of financial markets and asset prices within a restricted time interval following quantitative easing events based on the assumption that the main explanation for any movements in such windows is the quantitative easing event itself. It’s a different way of holding all other things equal. The following section provides a review of the literature and some of the key results obtained.

5. The event study literature

It’s usual to expect that credible policy announcements will be factored into market prices when they are made rather than when the policy actions themselves actually occur because this is when participants will form their expectations and valuations (Gagnon et al., 2011). So,
in the existing literature, researchers have typically monitored the reactions of financial markets in relatively short time-periods following Bank of England announcements on quantitative easing.

5.1 Gilt Market Reactions

The policy’s impact is hard to quantify, but the consensus in the literature is that the Bank of England’s quantitative easing programme has had significant effects, at least on bond yields (Joyce et al., 2012). The difficulty in being precise about the impact of quantitative easing is reflected in a wide range of estimates across studies. In an early study, Meier (2009) estimated that gilt yields fell in the first round of asset purchases by at least 35-60 basis points. Meier (2009) supported the event study with a comparison of gilt yields and foreign government bond yields. In July 2009, UK 10-year gilt yields were 10 basis points below their early March levels, while 10-year government bond yields in the US, Germany and Switzerland had risen by between 25 to 50 basis points over the same period. Perhaps most notably, Joyce et al. (2011) estimated that the asset purchases in the first year of the programme had reduced gilt yields by just under 100 basis points. The study estimated that the reactions ranged between 55 and 120 basis points across the 5 to 25-year part of the yield curve (Joyce et al., 2011). Compare this with the observation made in the previous section that the total fall in gilt yields during the programme was roughly 200 basis points.

In contrast to these findings, however, some other studies have suggested a smaller impact. Glick and Leduc (2011) and Meaning and Zhu (2011) both estimate that the effect of quantitative easing was closer to 50 basis points. This difference is likely to be due to the fact that both elected to use a 1-day window, whereas Joyce et al. (2011) used a 2-day window, finding significant impacts in the second day after an announcement. Event study analysis of financial markets often involves a short intraday interval, but this was deemed inappropriate by most researchers, as markets would need longer to incorporate the news given the novelty of quantitative easing. Joyce and Tong (2012) examined high frequency, intraday data on individual gilts, allowing them to date the market reaction to the exact time that the announcements were made rather than from the close of business the day before. Both intraday studies estimated that the first round of quantitative easing had an impact consistent with Joyce et al. (2011) of around 100 basis points. However, they also found that throughout the programme, the gilt market took varying amounts of time to incorporate quantitative easing announcements into prices, reducing the validity of results via event studies.
This reflects that this method of analysis relies on the ability to attribute the market reaction to a single event independent of anything else. Evidently event study analysis is sensitive to the length of the window used. In this case, due to the severity of the global crisis, there was regularly other relevant news around the time of announcements driving government bond yields in different directions (Joyce, McLaren and Young, 2012). If the window is too short the full reaction may not be captured and if the window is too long then the reaction could be contaminated by other events (Joyce et al., 2012).

It’s also worth noting that event study analysis suggests that the policy’s main impact was after the first two announcements. According to several studies, the largest reactions occurred in response to the February 2009 Inflation Report and the March 2009 announcement on the launch of quantitative easing itself (Joyce et al., 2011; Joyce, McLaren and Young, 2012). It seems that the reaction to later announcements was noticeably smaller. Indeed, many reports suggest that the second phase of quantitative easing, from October 2011 to October 2012, had little to no impact (see Meaning and Zhu, 2011; Joyce, McLaren and Young, 2012; Martin and Milas, 2012). Joyce, McLaren and Young (2012) even go so far as to associate the £125 billion of gilts purchased between October 2011 and May 2012 with a small rise in medium to long-term gilt yields.

However, there is no conclusive evidence that quantitative easing became weaker over time, and it’s possible that later rounds of quantitative easing had an impact that was comparable to that of earlier rounds. First, gilt yields rose by less than international yields on announcement days between October 2011 and May 2012, which suggests that other economic news, including market-moving developments in the euro area, may have been affecting results during the event study windows (Joyce, McLaren and Young, 2012). Secondly, the fact that event studies show smaller reactions over time reveals a fundamental flaw of this method of analysis. It’s likely over time that the market gained a better understanding of the programme and was more able to accurately anticipate the relevant announcements (Joyce, McLaren and Young, 2012). In measuring the immediate reaction to announcements, event studies reveal the extent to which the news is different to expectations. They measure extent to which the event was a surprise, rather than the magnitude and impact of the event itself. It’s possible that correctly anticipated moves have an impact prior to announcement, and that news that is contrary to expectations has an impact on the day of the announcement that reverses one that happened previously, or that is itself reversed subsequently. There may also be moves on days when there was no announcement but one was expected.
This helps to explain why there was such a large movement in shorter-maturity bonds following the February 2009 Inflation Report as the original and mistaken market perception was that the asset purchases would be targeted at shorter-dated gilts (Joyce et al., 2011). Following the March 2009 announcement in which the details of the programme were revealed, the largest effect was seen in 15 to 20-year maturities (up to 80 basis points), reflecting a correction in the earlier expectations (Joyce et al., 2011). Joyce et al. (2011) note that the announcements that followed up until November 2009 in which the programme was expanded to £200 billion were widely anticipated, which explains why the event study approach finds little reaction in the gilt market. Joyce, McLaren and Young (2012) confirm that the phase of quantitative easing announced in October 2011 was also widely expected due to the deterioration in the medium-term outlook for UK inflation, and thus it is probable that much of the effect of quantitative easing had already been incorporated into gilt prices prior to the event study window.

5.2 Reaction of corporate bonds, UK equities and exchange rate markets

Assuming investors do not view money as a perfect substitute to gilts, we would expect to see a response in the prices of other assets, such as corporate bonds and equities, and also possibly the sterling exchange rate. Announcements about quantitative easing may also affect perceptions about the future performance of the economy, and thus future corporate earnings; and changes in gilts may have an effect on the rate at which investors discount future cash flows (Joyce et al., 2011). Therefore, it’s possible that quantitative easing will have had an impact on other asset prices and the exchange rate.

The literature is less clear about these other impacts. Joyce et al. (2011) estimate that the first six announcements outlined in Figure 1 caused a 70 basis point fall in Sterling investment grade corporate bond yields, and a 150 basis point fall in non-investment grade corporate bond yields. According to their study, the effective exchange rate index fell by 4%. However, the largest fall by far was following the February 2009 Inflation Report, which also indicated the likelihood of a further short rate cut, meaning that this fall may not be solely related to quantitative easing (Joyce, McLaren and Young, 2012).

Throughout the programme, there was no uniform response to quantitative easing news in equity prices according to the literature (Joyce et al., 2011; Joyce, McLaren and Young, 2012).
As is the case with gilt yields, event study analysis in the existing literature indicates that the second phase of quantitative easing, lasting a year from October 2011, had very little impact. Joyce, McLaren and Young (2012) find an incremental increase and decrease in investment and non-investment grade corporate yields, respectively. However, they emphasise that this is because news regarding quantitative easing was widely anticipated by the markets.

It is also noted by Joyce et al. (2011) that the method of observing an immediate reaction to announcements relating to quantitative easing is less suitable when examining the impact on other assets. This is because it may take time for market participants to fully rebalance their portfolios and, thus, for the full impact of quantitative easing to be incorporated into asset prices.

5.3 Evidence of transmission channels

The literature on this topic provides considerable evidence to show that the main impact from quantitative easing in the UK has come through the portfolio-rebalancing channel. This was expected by the Bank of England when they designed the policy as detailed in section 3.2. Much of the research uses interest rates from Overnight Index Swap (OIS) contracts to shed light on the significance of the role that the portfolio rebalancing channel plays compared to the role of short-term rate expectations. OIS are derivative contracts involving the exchange of a predefined fixed interest rate with one linked to a compounded overnight interest rate (Joyce et al., 2011). They are popular amongst financial institutions and considered to provide an accurate measure of default risk-free rates. If the main transmission mechanism of quantitative easing was through an impact on future expected short-term rates, one might expect to see a corresponding movement in OIS rates (Joyce, McLaren and Young, 2012).

Following the February 2009 and March 2009 announcements, OIS rates fell around half as much as gilt yields, indicating that some of the initial influence of the policy came through a change in short rate expectations. However, the move in gilt yields was much larger than that in OIS rates in the other event study windows (Joyce et al., 2011). The fact that gilt yields moved appreciably more over all the event windows has led to a belief in the literature that the portfolio balance effect is the most important channel of influence (see Joyce et al., 2011; and Joyce, McLaren and Young, 2012). Christensen and Rudebusch (2012) further support this interpretation by decomposing gilt yields into expectations about short-term interest rates and term premiums using dynamic term structure models. They found that declines in gilt yields reflected reduced term premiums more than lower expectations of future rates,
supporting the claim that the policy was successful in changing the relative supply of assets held by private investors.

5.4 Impact on the wider economy

It is outside this paper’s scope, but we might briefly mention that some studies have attempted to quantify the impact of the asset purchases on the macro-economy using model-based simulations. On the assumption that the asset purchase programme did indeed reduce gilt yields by 100 basis points, as suggested by the event study conducted by Joyce et al. (2011), the Bank of England (2014b) estimated that the peak cumulative impact on real GDP could have been around 2.5%. Whilst these findings are an interesting and illustrative contribution to the debate, they cannot yet be seen as definitive. Given that this policy has only been used in severe financial crises, when even conventional monetary policy has uncertain effects, it will probably be some time before we are able to determine the broader impact of quantitative easing (Breedon, Chadha and Waters, 2012).

5.5 Summary

For all its weaknesses, the event study literature does suggest that quantitative easing has had an effect on gilt yields ranging from around 50 basis points up to 100 basis points. Furthermore, there is evidence, albeit less strong, to conclude that the purchases may also have affected the prices of a wider range of assets. There is also evidence to suggest that the most influential transmission mechanism was the portfolio balance effect.

However, the event study research also suggests that the early rounds of asset purchase announcements may have had a larger impact than the later rounds. Earlier announcements may have had a larger effect on markets due to the novelty of the policy. This is a problem with the event study approach and one that this paper has sought to address through an expert survey as discussed in the following section.

6. The results of a qualitative survey

This section presents the findings from this paper’s own research. As discussed, it’s extremely difficult to accurately estimate the impact that quantitative easing has had on financial markets in the UK. The typical approach in the literature has been to use event study analysis, which, as explained in the above sections, is limited due to the fact that it measures the extent to which the market has expected the news regarding quantitative easing, rather than what the news itself is. In order to overcome this problem with the previous literature, this paper
offers a survey into the attitudes and opinions of a small sample of individuals in an attempt to provide an insight into what the experts believe the impact of quantitative easing to have been in the UK.

The weaknesses of this approach are that the sample was not designed to be representative of any wider population and the responses are in many cases qualitative. On the other hand, the strength is that a reasonable number of well-placed and well-informed respondents are, by definition, being invited to assess the impact of quantitative easing separately and independently of the various other factors and influences at work at the time. As we have seen this is extremely difficult, but this paper believes that this means of approaching the task is complementary to the others, and we are particularly interested where we find similarities between our conclusions and those made by other methods of analysis.

The sample of 35 experts included a range of commentators, financial market practitioners and former policy-makers. Some current policy-makers declined to participate. The sample was intended to contain a cross-section of people who were likely to have an opinion on the issues of concern. The survey asked them to evaluate six key aspects of quantitative easing. The questions were as follows. How important is each channel of influence? Respondents were asked to rate the importance of money supply, the exchange rate, short-rate expectations, the gilt market and other asset markets on a scale of one to five. They were also asked whether the stock or the flow of gilt purchases was the more appropriate parameter in terms of measuring the impact of the programme. The survey asked what was the estimated impact on the 10-year gilt yield, and whether this adequately summarised the impact of the programme more generally. The survey asked which other markets and asset prices were likely to have been affected by the policy and/or any fall in gilt yields. The options were the UK equity market, the sterling exchange rate, UK residential property and alternative asset markets. Lastly, the respondents were asked what is the most likely exit strategy. There was a high measure of agreement on some of the questions, but less so on others. For charts illustrating the responses to each question, the reader is referred to Appendix 5.

6.1 Importance of transmission channels

There was a lack of a consensus in the sample on the importance of the money supply as a transmission channel, with responses ranging across the scale. In part, this may reflect a disinterest in, and unfamiliarity with, the money supply as a concept. More respondents thought that the impact on expectations or the ‘signalling’ channel was important. However,
the lack of consensus was evident again, with almost a quarter of respondents deeming this transmission channel completely insignificant. The clear majority of respondents thought that the policy's impact on the gilt market and other asset markets was important or very important. This would support the view discussed in the literature that the portfolio rebalancing effect played a large role in the UK programme. Although there were mixed responses on the importance of the exchange rate, it’s worth noting that the majority of respondents saw some of the policy's effect come through the exchange rate. This is an interesting finding given that policy-makers rarely speak of this channel of transmission due to the political consequences of a policy that will depreciate the exchange rate.

From the responses to this question, we would suggest that there is clear agreement that quantitative easing had a strong effect through the gilt market and other markets. Thus, the results from the survey suggest that the portfolio rebalancing effect was most powerful, complementing the existing literature. There is some recognition of other channels of influence, but no real consensus concerning their importance.

6.2 Importance of stock/flow of gilt purchases

The majority (just over 60%) of respondents suggested that the flow of asset purchases mattered more than the stock. The Bank of England has typically suggested the opposite (Dale, 2010). The survey results suggest that the largest impact of quantitative easing would have been while the purchases were being made, with a smaller ongoing impact when, as now, that was no longer the case.

6.3 Impact on gilts and other markets

The responses to the question on the size of the impact on gilt yields were evenly split between the ‘up to 50 basis points’ category and the ‘between 50 and 100 basis points’ category. On average, then, we might suggest that the respondents view the impact to have been around 50 basis points. Interestingly, this is broadly consistent with some of the event study results summarised above and, again, can be compared with the observation made in a previous section that the total fall in gilt yields during the programme was roughly 200 basis points.

The sample was also clear that the impact on 10-year gilt yields doesn’t summarise the full impact of the programme. This, again, supports the view that portfolio rebalancing was important. The impact was felt most immediately and most strongly in the gilt market, but
other markets have seen secondary or subsequent impacts. In particular, a considerable number of respondents pointed to a strong effect on UK equities and, less so, on the exchange rate. It’s worth noting that much of the previous research on the topic fails to identify whether the equity market or the exchange rate was significantly affected, so the responses to this question make a useful contribution. There were also some who thought that there had also been an effect on property and alternatives. In addition, the sample was almost unanimous that the exit strategy would consist of raising short rates rather than making asset sales.

6.4 Summary

Overall, while it’s difficult to draw strong conclusions from the survey, we would highlight some key findings. First, there is a clear consensus that the impact on gilts and other assets was important to the working of the policy. Second, there is a consensus that the impact of the policy is not adequately summarised by its impact on gilt yields. Both these points are consistent with the importance of the portfolio balance effect, which is also supported by market pricing and some of the literature on the topic. Third, the survey would suggest that the impact of quantitative easing on 10-year gilt yields is, on average, believed to have been around 50 basis points. This is consistent with some of the event studies, and suggests that others may have overestimated the impact of the policy. However, a weakness of the survey is that it cannot produce a precise estimate from the results; it isn’t possible to estimate a more accurate range than 0 to 100 basis points. Finally, there is significant belief that the policy had a strong impact on the UK equity market and, to a lesser extent, the sterling exchange rate.

The fact that a sample of experts agreed on a number of points but not on others, most notably the question on transmission channels, is also an important revelation of the survey. Whilst there is a general consensus that the policy had the intended impact, there is less agreement on how and why that came about. Further research is therefore encouraged in order to further our understanding.

7. Conclusion

The extraordinary combination of financial crisis and recession presented a major challenge to policy-makers in governments and central banks around the world. It’s no surprise, then, that the policy response was both unprecedented and unconventional. It is, however, a relief that it appears to have been successful. It’s clear that gilt yields fell significantly during the period in which quantitative easing took place. It’s important to remain cautious in attributing
that to the programme itself. To state the obvious, observations made during the period in which asset purchases were made cannot necessarily be attributed to the asset purchases themselves.

Despite its weaknesses, the event study literature does suggest that quantitative easing has had an effect on gilt yields that could range from around 50 basis points up to 100 basis points. Furthermore, there is evidence, albeit less strong, to conclude that the purchases may also have affected the prices of a wider range of assets. There is also evidence to suggest that the most influential transmission mechanism was the portfolio balance effect.

The contribution of this paper relative to previous work on the financial market impact of quantitative easing has been to conduct a qualitative survey of market experts. The results from the survey show a clear consensus that the impact on gilts and other assets was important to the working of the policy, and, also, that the impact of the policy is not adequately summarised by its impact on gilt yields. Both points are consistent with the importance of the portfolio balance effect, and thus complement and validate the literature. The results would also suggest that the impact on gilt yields is, on average, believed to have been around 50 basis points. Studies have tended to use the same event study methodology, so our qualitative survey is a useful contribution to the literature as it widens the evidence base. Finally, there is significant belief that the policy had a strong impact on the UK equity market and, to a lesser extent, the sterling exchange rate. This is a noteworthy contribution that many previous studies have failed to discover.

Our provisional conclusion is that quantitative easing has played a significant part in easing monetary policy in the UK in much the way that was intended and expected. In so doing, we believe it will also have contributed to the stabilisation of the UK economy after one of the most difficult periods in recent economic history. However, it remains difficult to accurately estimate the size of the effect that the policy has had. This is a challenge that needs to be tackled by future research in order to increase our understanding of quantitative easing. It’s possible that our survey-based approach can be extended to include more precise estimations from industry experts. The lack of a consensus on the size of the policy’s impact on intermediate targets makes it harder to determine the extent to which quantitative easing can influence spending and inflation. This is ultimately what the policy will be judged on.
Bibliography


Appendices

Appendix 1

Chart 8: UK GDP growth (quarter on same quarter previous year, %)

Source(s): this paper’s own illustration of data taken from the Office for National Statistics, April 2015

Appendix 2

Chart 9: UK unemployment (%)

Source(s): this paper’s own illustration of data taken from the Office for National Statistics, April 2015
Appendix 3

Chart 10: UK CPI inflation (%)

Source(s): this paper’s own illustration of data taken from the Office for National Statistics, April 2015
Appendix 4

The survey

1. In evaluating the impact of QE, how important are the following channels of influence – please award marks out of 5 to options A, B, C and D. (5 is very important and 1 is very unimportant)

A. Impact on the money supply
B. Impact on short rate expectations
C. Impact on gilt and other asset prices
D. Impact on the exchange rate

2. What is more appropriate in measuring the impact and extent of QE? Please circle ‘a’ or ‘b’.

A. The stock of gilt purchases
B. The flow of gilt purchases

3. Is the impact of the programme on financial markets adequately summarised by the impact on 10-year gilt yields? Please circle ‘A’ or ‘B’.

A. Yes
B. No

4. What do you estimate the impact on 10-year yields to have been? Please circle A, B, C, or D.

A. Negligible
B. Up to 50 basis points
C. Between 50 and 100 basis points
D. Over 100 basis points

5. What other markets and asset prices are likely to have been affected by QE and/or any fall in gilt yields - please award marks out of 5 to options A, B, C and D. (5 is significantly affected and 1 is not affected)

A. The UK equity market
B. The sterling exchange rate
C. UK residential property
D. Alternative asset markets

6. What is the most likely exit strategy? Please circle ‘A’ or ‘B’.

A. Raising short rates without selling gilts back
B. Selling gilts back without raising short rates
Appendix 5

Charts illustrating the results from this paper’s qualitative survey

Question 1

Chart 11: Channels of influence: money supply

Chart 12: Channels of influence: expectations
==Chart 13: Channels of influence: gilts and other asset markets

Chart 14: Channels of influence: exchange rate
Question 2

Chart 15: Quantifying QE: stock or flow?

Quantifying QE: stock or flow?

Question 3

Chart 16: Quantifying QE: do gilt yields summarise impact?

Quantifying QE: do gilt yields summarise impact?
Question 4

**Chart 17: Quantifying QE: impact on 10 year gilt yields**

Quantifying QE: impact on 10 year gilt yields

![](chart17.png)

Question 5

**Chart 18: Other asset markets: UK equities**

Other asset markets: UK equities

![](chart18.png)
Chart 19: Other asset markets: exchange rate

Chart 20: Other asset markets: property
Question 6

Chart 21: Other asset markets: alternatives

Chart 22: Exit strategy: raise rates or sell gilts
The Impact of the ‘One-Child Policy’ on China’s Aggregate Household Savings – An Econometric Analysis

Lele Ding

The purpose of this dissertation is to explore the impact of the One-Child Policy on China’s aggregate household savings. There are four proxies for the One-Child Policy: the total fertility rate, annual population growth rate, age structure, and gender ratio. The purpose of this dissertation thus has been achieved by investigating the statistical relationships between these four demographic factors and China’s aggregate household savings, in light of the theory of Life-Cycle Saving. An ordinary least squares regression model has been conducted based on the cross-sectional annual data from the National Bureau of Statistics of China and the World Bank for the period of 1980-2012. The results of multiple regressions illustrate that the One-Child Policy has a statistically significant relationship with China’s aggregate household savings.
1. Introduction

China’s One-Child Policy (OCP) is unique in the history of the world (Choukhmane et al., 2014; Greenhalgh, 2008). Before the implementation of the OCP from the 1950s through the 1960s, the total fertility rate (TFR) in Mainland China was approximately 6 children born per woman (The World Bank, 2015). Since the enforcement of the OCP by the Chinese government in 1980 (Greenhalgh, 2008), the TFR in China has continually declined to roughly 1.66 children per woman of childbearing age in 2012 (The World Bank, 2015). Additionally, during the same time period, China’s aggregate household savings (CAHS) has experienced an opposite tendency and by 2012 it achieved approximately 39,955.1 billion Chinese Yuan (CNY), an amount that was almost 1,000 times the level before employing the OCP (NBSC, 2013a, citied in the People’s Bank of China, 2013, p. 1). Thus, the phenomenon that the rising CAHS coincides with the time, in which the OCP was implemented, forms the main motivation of this dissertation. This topic is important for Chinese policy makers in order to understand the role of demographic policy, such as the OCP, in planning for the mobilisation of CAHS and the formulation of other relevant economic policies.

The dissertation aims to conduct a multiple regression analysis of a cross-sectional study in investigating the potential relationship between the OCP and CAHS - examining the effects of the OCP-related variables on CAHS. The term CAHS in this dissertation is measured as the aggregate balance of deposits of Chinese households in banks and other financial institutions (NBSC, 2013a). In order to achieve the purpose of this study, there are four objectives that will be explored. The first objective is to present China’s demographic characteristics and the stylised facts of CAHS. Secondly, the existing literature resources on this topic will be discussed to identify the applicable theory of Life-Cycle Saving (LCS) and to help determine the likely variables for the regression models in this dissertation. Based on these pieces of information, the third objective is to set up my hypotheses and then develop an empirical investigation.

The rest of the dissertation is organised as follows. Section 2 provides background information on the evolution of the OCP and its accompanying impacts on the Chinese population, as well as the development of CAHS. Section 3 critically discusses the existing literature concerning the impact of the OCP on CAHS, from both theoretical and empirical perspectives. Section 4 provides a framework for the empirical investigation, which encompasses the methodology, results, discussion and limitations. Additionally, the methodology section involves eleven stages. The first stage simply introduces the annual data from the National Bureau of Statistics of China (NBSC) and the World Bank between 1980 and 2012. The second stage discusses the initial variables that are suggested by
the former literature. In the third stage, scatterplots are used to inspect whether a data pattern is linear or nonlinear, aiming to select an appropriate functional form. The next two stages identify the variables and the model applied in this study, respectively. Their validity will be verified by a set of tests, which are included in the remaining stages. Along with relevant theories and literature, the empirical results indicate that the OCP-related variables are jointly significantly in estimating the impact of the OCP on CAHS. The final section highlights limitations and areas for improvement in further studies. The final chapter concludes that the OCP has detectable effects on CAHS.

2. Background

This chapter represents the evolution of the OCP and the corresponding demographic transformation, which is accompanied by the changes in CAHS. Figure 1 demonstrates the trends of the TFR and the annual population growth rate (APGR) in China from 1964 to 2012. Figure 2 displays the growth of CAHS during the same period.

Figure 1: The Total Fertility Rate and the Annual Population Growth Rate in China, 1964-2012

2.1. Prior to the Implementing the OCP (1949-1979)

Since the establishment of the People’s Republic of China in 1949, the first Chairman, Mao Zedong, encouraged Chinese people to have more children in order to accumulate more human labour (Greenhalgh, 2008). As a result, China’s population has dramatically increased from approximately 5.4 billion in 1949 to 8 billion in 1969 (NBSC, 2013b), with an average of 6 children per family (The World Bank, 2015). China accounted for roughly 20 per cent of the world population by the end of 1969 (The World Bank, 2014b); however, it only covered 7.2 percent of the world’s landmass by the end of 1969 (The World Bank, 2014a). This can be argued as unsustainable because the population could grow beyond the resources of the country (Greenhalgh, 2008). Hence, in the 1970s, the Chinese government has tried to control Chinese population by the less mandatory approach, such as family planning. It had advocated each family to have later and fewer children (Greenhalgh, 2008). China’s TFR had thus declined from 5.5 births in 1970 to 2.8 births per woman in 1979, approximately (see Figure 1) (The World Bank, 2015). Additionally, the APGR in China dramatically reduced to only 1.3 per cent by 1979 (The World Bank, 2013). Under these conditions, China’s second-generation leader Deng Xiaoping believed that forcibly restricting Chinese population growth would benefit greater economic prosperity and social development (Potts, 2006). Therefore, the OCP was announced in the late 1979 and formally instituted nationwide by the Central Committee of the Chinese Communist Party on September 1980 (Greenhalgh, 2008). Moreover, CAHS maintained
at the relatively low level from 1949 to 1979 (see Figure 2), grew at a rate of 15 per cent annually (NBSC, 2013a, cited in the People's Bank of China, 2013, p. 1).

2.2. After the Introduction of the OCP (1980-2012)

The OCP is a population control policy that aims to alleviate social, economic and environmental consequences of overpopulation (Kane and Choi, 1999). The policy stipulated that Chinese couples of childbearing-age could only have one child, which applied to the majority of the ethnic Han Chinese (Greenhalgh, 2008). According to the 2010 Chinese Census, Han Chinese accounts for 91.5 percent of the total population in Mainland China (Howden and Zhou, 2014). The OCP has been more strictly enforced in urban areas than in rural areas. Rural couples were allowed to have two children if their first-born child was female (Greenhalgh, 2008).

There are four potential demographic impacts induced by the OCP. Firstly, the TFR decreased to an average of 1.66 births per woman in 2012 (The World Bank, 2015) and it was below the replacement fertility rate of roughly 2.1 births per woman (Gu and Li, 2010). In addition, the APGR dropped to only roughly 0.487 per cent by 2012 (see Figure 1) (The World Bank, 2013). These changes had significantly accelerated China’s age structural transitions (Gu and Li, 2010). As can be seen from Figure 3, the proportion of the Chinese population that is aged 0-14 was gradually reduced by roughly 17.4 per cent from 1980 to 2012 (The World Bank, 2014c). The percentage of this age group in 2012 was twofold below the value in the year before introducing the OCP (The World Bank, 2014c). Additionally, the shares of both age groups of 15 to 64, 65 and over age have increased by more than 13 and 3.5 per cent, respectively (The World Bank, 2014d; 2014e). The final demographic effect refers to the imbalanced gender ratio, which has become skewed toward males in China after enforcing the OCP (see Figure 4) (NBSC, 2013b; Li et al., 2011). When the majority of Han Chinese families was strictly restricted to have one child, male children have been preferred and female children have become highly undesirable in traditional China’s context (Li et al., 2011). This thus had led to an increase in abortions of female foetuses, particularly in China’s rural areas. Furthermore, returning to Figure 2, CAHS has experienced a significant growth of 44720.58 billion CNY in the OCP period from 1980 to 2012. This increase was much larger than in the 17-year period between 1962 and 1979 when CAHS increased by 24.96 billion CNY (NBSC, 2013a, cited in the People’s Bank of China, 2013, p. 1).
Figure 3: Chinese Population by broad Age Groups (1964-2012)

Sources: based on the World Bank, 2014c, p.1; 2014d, p.1; 2014e, p.1

Figure 4: Stacked Line Chart of Comparing Male and Female Population in China (1964-2012)

Source(s): based on NSBC, 2013b, p. 1.
2.3. The Easing of the OCP (2013-Present)

Chinese officials announced the aforementioned four demographic problems have been largely caused by the OCP (The Economist, 2013). In response to this, on March 2013, the Chinese government decided to slightly ease the OCP to a ‘two child policy’ in many Chinese urban districts. The new policy has adjusted the OCP by transforming from one child to two children, so long as either of the parents is an only child (The Economist, 2013). The relevant data and information of the new policy have not been released, so that its effects on China’s demographic problems and CAHS need to be further investigated.

3. Literature Review

The existing academic literature that investigates the relationship between the OCP and CAHS is relatively scarce. These studies are limited because they examine this question by isolating each potential demographic impacts engendered by the OCP. However, these studies have helped accentuate applicable theories and identified four key factors to be the indicators of the OCP. Therefore, this can be advantageous to the empirical analysis in this dissertation. The four factors are: changes in the TFR (Horioka and Wan, 2007; Modigliani and Cao, 2004), the declining annual population growth (Fan and Zhu, 2012; Ito and Rose, 2010; Horioka and Wan, 2007; Cook, 2006), age structure transition (Choukhmane et al., 2014; Zhu et al., 2014) and imbalanced gender ratio (Bulte et al., 2011; Li et al., 2011; Ebenstein, 2010; Zhu et al., 2009; Wei and Zhang, 2008; Zeng et al., 1993).

According to the relevant literature, the theory that underpins the topic of this dissertation is Life-Cycle Saving (LCS), which is commonly used to describe people’s overall saving behaviour (Modigliani, 1986; Ando and Modigliani, 1963). There are two main assumptions behind the theory. Firstly, there is a relatively smoother consumption pattern over people’s entire life spans (see Figure 5) (Modigliani and Cao, 2004). Secondly, people’s lifetime income is likely to differ systematically, following a hump-shaped pattern. Specifically, the entire lifetime can be divided into three periods: the youth (children), the middle-aged (parents) and the old-aged (retired) (Modigliani and Cao, 2004; Ando and Modigliani, 1963). Children are expected to be dissaving in their first period, they will rely on their parents’ supports, including childcare costs and education investments (Ando and Modigliani, 1963). In return, children would offer part of their future wage income to their parents’ retirement consumptions when they enter the workforce in the next period. It assumed that only the middle-aged people would work for wage incomes and be responsible for their children’s subsistence consumptions, educational investments and their parents’ retirement consumptions.
(Modigliani and Cao, 2004). They would also accumulate more capital to ensure their own retirement consumption (Ando and Modigliani, 1963). In the final period, the retired people’s incomes fall and they would negatively contribute to aggregate savings (Horioka and Wan, 2007; Modigliani and Cao, 2004).

**Figure 5: A Life-Cycle Saving Model**

![Diagram of a life-cycle saving model](image)

Source(s): based on Modigliani and Cao, 2004; Ando and Modigliani, 1963.

First of all, Becker and his associates (Blake, 1989; 1981; Becker and Tomes, 1976; Becker and Lewis, 1973) develop the theory that there would be a quantity-quality trade-off of children within a family given the fertility restriction. Many other scholars verify that this theory can be used to partly explain CAHS given the OCP (Choukhmane et al., 2014; Zhu et al., 2014; Angrist et al., 2005; Ginther and Pollak, 2003; Behrman et al., 1994). This result is also consistent with Rosenzweig and Zhang (2009) and Li et al. (2008). Additionally, they utilised the Chinese Census and discovered a negative correlation between the TFR and expenditure on the education of children. This encourages families to save more (Li et al., 2008). However, Ahn et al. (1994) and Blake (1981) have not come to the same conclusion. This is because they simply regard the TFR as an exogenous variable and ignore the fact that children’s development is also affected by their parents (Choukhmane et al., 2014; Zhu et al., 2014; Rosenzweig and Zhang, 2009).

Turning to the factor of APGR, Cook (2006) argues that population growth would increase the proportion of the working-age population in society, which is positively related to the aggregate household savings and vice versa. Moreover, it is widely recognised that the decreased APGR in
China has always been associated with the changes in age structure of the population (Ito and Rose, 2010). In an analysis of the effects of the age structure transition on CAHS, Cook (2006) reinforces the view that household’s propensity to save depends on people within an age group number of each age groups. Following the theory of LCS, savings generated by the middle-aged people are more than the dissavings generated by the youth and the elderly (Cook, 2006). However, his argument has not been assessed by quantitative methods.

In China’s context, the measurements of the age structure employed in most studies are the proportion of three age groups (0-14 years, 15-64 years, 65 years and over) (Choukhmane et al., 2014; Zhu et al., 2014; Liao, 2013; Ma and Yi, 2010). Nevertheless, based on the dynamic panel data between 1995 and 2004 from China’s household survey, Horioka and Wan (2007) maintain that there is no significant impact of the demographic composition on CAHS. This argument can be critiqued as inconclusive, as they chose an incomplete time period of 1995-2004. As a result, this would affect statistical significance of estimators. In the meantime, the problem of multicollinearity occurred when putting the young and the elderly aged dependency ratios in the same regression. Therefore, the accuracy of estimation and prediction proposed by Horioka and Wan (2007) is questionable.

Moreover, based on the general equilibrium overlapping generations model, China’s Census databases, as well as China Health and Retirement Longitudinal Study, Choukhmane et al. (2014) and Zhu et al. (2014) conclude that the reduction in overall expenditures owing to fewer young people (ages 0-14) has contributed to higher CAHS. However, Zhu (2011), Kelley and Schmidt (1995) and Weil (1994) explain this from a different perspective. The decline in the young population would have the consequence of a fall in the middle-aged people. This, however, would be negatively related to aggregate household savings (Zhu, 2011; Kelley and Schmidt, 1995; Weil, 1994).

Additionally, there are statistically significant results suggesting that the middle-aged population in China is motivated to accumulate more physical and human capital in order to be responsible for their next generation and their retired parents (Choukhmane et al., 2014; Zhu et al., 2014; Banerjee et al., 2010; Wang, 2009). As for the elderly population in China, Zhu et al. (2014), Modigliani and Cao (2004) and Kraay (2000) state that they tend to save more with a lower fertility rate. The result is also supported by Ito and Rose (2010), who concluded that private savings of the elderly people in China are largely from public pensions and family’s financial support and both also positively contributed to CAHS. What is more, there is another interpretation that the elderly would
accumulate more financial wealth in expectation of lower support from their children, given the fertility restriction. In other words, a rise in the elderly population can have the effect to cause aggregate savings increasing, not decreasing (Choukhmane et al., 2014; Zhu et al., 2014; Wang, 2009).

Finally, the male-biased sex ratio in China has partially resulted from male preference and gender-selection technologies, which was mainly attributed to the enforcement of the OCP in 1980 (Bulte et al., 2011; Li et al., 2011; Ebenstein, 2010; Zhu et al., 2009; Zeng et al., 1993). Li et al. (2011) make use of the 1990, 2000, and 2005 Chinese Census data and a difference-in-difference estimator to predict that the OCP has led to 7.0 extra males per 100 females. Bulte et al. (2011) and Wei and Zhang (2009) conclude that the imbalanced sex ratio in China could account for approximately half of the actual increase in CAHS between 1990 and 2007. Their findings were authenticated based on Chinese regional and household-level survey data. As they explained, Chinese households with a son tend to increase their savings in order to improve their son’s competitiveness in marriage markets. In addition to that, households with a daughter in both urban and rural areas do not reduce their savings, probably because their parents attempt to improve their daughter’s bargaining power after marriage. This is held in studies by Li et al. (2011) and Wei and Zhang (2008) who conclude that the male-biased sex ratio in China has a statistically significant positive impact on CAHS. By the way of comparing savings survey data in households with sons versus those with daughters, their results illustrate that households with sons tend to increase their savings more than households with daughters on average, and this becomes more apparent where households in a region with a more skewed sex ratio (Li et al., 2011; Wei and Zhang, 2011; 2008).

On the contrary, other empirical studies suggest that aggregate household savings would decline in response to a male-biased sex ratio (Griskevicius et al., 2012; Weir et al., 2011; Kvarnemo and Anhesjo, 1996; Taylor and Bulmer, 1980). Due to the increasing intensity of the competition between the same-sex, men would be motivated to increase their consumptions during their courtship in order to signal their attractiveness. As a consequence, aggregate household savings would decline. Nonetheless, this finding has not been verified in the case of China given the OCP.

Therefore, this study will aim to explore whether there is empirical evidence to show that the introduction of the OCP can affect CAHS, in light of the theory of LCS. Furthermore, the empirical model in this dissertation will only measure the impacts of the four factors (the proxies for the OCP) on CAHS. Importantly, there are three main research gaps derived from the aforementioned
literature, which will be improved in this dissertation. Firstly, there are no studies examining the impact of the OCP on CAHS – grouping together the OCP-related factors to simultaneously assess their impacts on CAHS. Secondly, there will be a larger sample size. Using a 33-year period (1980–2012) allows an appropriate length of time for the evaluation of the policy effect: the OCP has been carried out from 1980 to 2012. The third contribution is to use the number of male population as the indicator of the male-biased gender ratio, in order to avoid the problem of collinearity in the later empirical analysis.

4. Empirical Investigation

4.1. Methodology

The empirical strategy in this dissertation is to conduct an econometric study with multiple regressions based on the cross-sectional data. Moreover, the coefficients of these regressions will be estimated using ordinary least squares (OLS) methods. There are three reasons for choosing this methodology. Firstly, an econometric study is a quantitative analysis of an actual economic phenomenon based on theories and observations (Stock and Watson, 2012). Secondly, multiple regression analysis allows investigating the joint effect of all targeted explanatory variables on the dependent variable, which is more informative in predicting the impact of the OCP on CAHS. Thirdly, the purpose of a cross-sectional study is to investigate the relationships between interest variables and this is in accord with the title of this dissertation (Stock and Watson, 2012). The following eleven stages will be explored to help generate more valid empirical results.

4.1.1. The Data

The annual data used in this study are derived from the National Bureau of Statistics of China (NBSC) and the World Bank throughout the time period 1980-2012. The set of data can be classified into the cross-sectional data, since it consists of multiple entities observed at a particular time period (Stock and Watson, 2012). The annual data on China’s aggregate household savings, the percentage of Chinese population that is male, as well as China’s total enrollment in the higher education are extracted from the NBSC. Moreover, China’s TFR, annual population growth rate, the Chinese population aged between 0-14, 15-64, 65 and over are taken from the World Bank database. The data employed are contained in the appendices.

It is also important that the data for the years of 1982, 1990, 2000 and 2010 from the NBSC are the Census year estimates. Except that, other data sets from the NBSC are estimates from the annual
national sample survey conducted by the Department of Population and Employment Statistics (NBSC, 2013b). Compared with other informal survey data, the Chinese official data seems to be more representative and reliable. The data from the NBSC are inconsistent with the data published in the World Bank. This, according to Stock and Watson (2012), would not introduce bias in the further empirical analysis, because the data are missing on the value of variables. Overall, it can be suggested that both data sources have relatively higher levels of accessibility and usability.

4.1.2. Initial Variables

A core set of initial variables has been selected on the basis of theoretical reasoning and suggestions from the literature reviews. The dependent variable is CAHS. Additionally, there are four independent variables, which are the TFR, annual population growth, demographic structure and gender ratio (Choukhmane et al., 2014; Zhu et al., 2014; Li et al., 2011; Ebenstein, 2010; Wei and Zhang, 2011; 2008; Zeng et al., 1993). Both the dependent and independent variables are measured at the continuous level. In this dissertation, the demographic structure factor is represented by three different age categories that are 0-14, 15-64 and over 64 years old (Choukhmane et al., 2014; Zhu et al., 2014; Liao, 2013). The number of male population is the proxy for the male-skewed sex ratio. Table 1 demonstrates the symbols and measure unit of all initial variables. Notably, an education variable (substituted by the total enrolment in the higher education) highlighted in the literature is also included in the regression analysis, because it can help the further discussion.

Table 1: Summary of Initial Variables

<table>
<thead>
<tr>
<th>Initial Variables</th>
<th>Symbols</th>
<th>Unit of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>China’s aggregate household savings</td>
<td>CAHS</td>
<td>Billion Chinese Yuan</td>
</tr>
<tr>
<td>Total fertility rate</td>
<td>TFR</td>
<td>Births per woman</td>
</tr>
<tr>
<td>Annual population growth</td>
<td>APG</td>
<td>Percentage</td>
</tr>
<tr>
<td>The percentage of male population</td>
<td>POMP</td>
<td>Percentage</td>
</tr>
<tr>
<td>Population aged 0 to 14</td>
<td>PA014</td>
<td>10,000 Persons</td>
</tr>
<tr>
<td>Population aged 15 to 64</td>
<td>PA1564</td>
<td>10,000 Persons</td>
</tr>
<tr>
<td>Population aged 65 and above</td>
<td>PAOVER64</td>
<td>10,000 Persons</td>
</tr>
<tr>
<td>The total enrolment in the higher education</td>
<td>TEHE</td>
<td>10,000 Persons</td>
</tr>
</tbody>
</table>

4.1.3. Function Form Specification

Before setting up an estimated regression model, it is helpful to examine a scatter plot of the data in order to visually see, which regression function is most likely going to be a good fit in this
dissertation. According to Stock and Watson (2012), if the relationship between the independent and dependent variables in scatterplots is non-linear, the general strategies are modifying the data and modeling nonlinear functions. In Figure 6, the scatterplots on the left side indicate linear-linear functions, and the log-transformed scatterplots are displayed on the right side. It can be seen that the right-side figures provide a better fit to the data and have approximately linear patterns. Thus, using logarithmic functions can make correlations between the independent and dependent variables more interpretable, based on a percentage basis. Apart from that, there are two additional advantages of using logarithmic functions. The first is that they do not have to be concerned about the units of measurement (Stock and Watson, 2012). Secondly, they can be a remedy for skewed data due to the presence of outliers (see Figure 6).

**Figure 6**: Scatterplots of Linearity between the Dependent Variable and Each of the Independent Variables

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6.1a)</td>
<td>CAHS and TFR</td>
</tr>
<tr>
<td>(6.1b)</td>
<td>ln(CAHS) and TFR</td>
</tr>
<tr>
<td>(6.2a)</td>
<td>CAHS and APG</td>
</tr>
<tr>
<td>(6.2b)</td>
<td>ln(CAHS) and APG</td>
</tr>
<tr>
<td>(6.3a)</td>
<td>CAHS and POMP</td>
</tr>
<tr>
<td>(6.3b)</td>
<td>ln(CAHS) and POMP</td>
</tr>
</tbody>
</table>
4.1.4. Variable Identification

Having selected relatively appropriate regression functions, the new variables used in the regression model in this dissertation will be presented in this section. The interpretations of all new variables are in Table 2.
Table 2: Summary of Key Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Symbols</th>
<th>Unit of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The logarithm of China’s aggregate household</td>
<td>ln(CAHS)</td>
<td>-</td>
</tr>
<tr>
<td>savings</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Independent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total fertility rate</td>
<td>TFR</td>
<td>Births per woman</td>
</tr>
<tr>
<td>Annual population growth</td>
<td>APG</td>
<td>Percentage</td>
</tr>
<tr>
<td>The percentage of male population</td>
<td>POMP</td>
<td>Percentage</td>
</tr>
<tr>
<td>The logarithm of population aged 0 to 14</td>
<td>ln(PA014)</td>
<td>-</td>
</tr>
<tr>
<td>The logarithm of population aged 15 to 64</td>
<td>ln(PA1564)</td>
<td>-</td>
</tr>
<tr>
<td>The logarithm of population aged 65 and above</td>
<td>ln(PAOVER64)</td>
<td>-</td>
</tr>
<tr>
<td>The logarithm of the total enrolment in the</td>
<td>ln(TEHE)</td>
<td>-</td>
</tr>
<tr>
<td>higher education</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additionally, Table 3 illustrates the descriptive statistics of these variables, including mean, standard deviation (SD) and extremum. The mean value is used to describe central tendency of the data set (Stock and Watson, 2012). The SD measures the dispersion in the distribution of a variable (Stock and Watson, 2012). For example, the SD of POMP is close to 0, which indicates that the data points tend to be very close to the mean of the set. While the variable POP014 has a relatively higher SD with approximately 5.27, implying the data set has more variability than expected. Thus, the measure of SD could be a useful estimate in the further analysis. The range of variables is indicated by the minimum and maximum (Stock and Watson, 2012).

Table 3: Summary of Descriptive Statistics

<table>
<thead>
<tr>
<th>Variables</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(CAHS)</td>
<td>33</td>
<td>9.96626</td>
<td>2.112902</td>
<td>5.980909</td>
<td>12.8981</td>
</tr>
<tr>
<td>TFR</td>
<td>33</td>
<td>2.031727</td>
<td>0.5260668</td>
<td>1.51</td>
<td>2.826</td>
</tr>
<tr>
<td>APG</td>
<td>33</td>
<td>1.006378</td>
<td>0.3929961</td>
<td>0.479151</td>
<td>1.610071</td>
</tr>
<tr>
<td>POMP</td>
<td>33</td>
<td>51.90987</td>
<td>0.3217522</td>
<td>51.35361</td>
<td>52.50872</td>
</tr>
<tr>
<td>ln(PA014)</td>
<td>33</td>
<td>10.33186</td>
<td>0.1285476</td>
<td>10.09463</td>
<td>10.45553</td>
</tr>
<tr>
<td>ln(PA1564)</td>
<td>33</td>
<td>11.28529</td>
<td>0.159795</td>
<td>10.97519</td>
<td>11.50343</td>
</tr>
</tbody>
</table>
4.1.5. Model Specification and Hypotheses

From what has been discussed above, the estimated non-linear regression model below is therefore preferred in this study (see Equation 1). Furthermore, the coefficients in this model are estimated using OLS and they are denoted $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6$, and $\hat{\beta}_7$. The OLS error term $\epsilon_1$ contains all the other potential factors besides the existing independent variables that determine the value of the dependent variables (Stock and Watson, 2012). The theory of quantity-quality trade-off of children (Blake, 1989; 1981) suggests that the coefficient of TFR is likely to be negative. Additionally, according to the theory of LCS (Modigliani, 1986; Ando and Modigliani, 1963), there are hypotheses: $\hat{\beta}_6 < 0$ and $\hat{\beta}_7 < 0$. While the sign of $\hat{\beta}_6$ is supposed to be positive. However, due to the lack of theoretical consensus on the effects of APG and POMP on CAHS, their corresponding signs of $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ cannot be determined a priori at this stage.

Equation 1: The Regression Model

$$
\ln(CAHS) = \hat{\beta}_0 + \hat{\beta}_1TFR + \hat{\beta}_2APG + \hat{\beta}_3POMP + \hat{\beta}_5\ln PA014 + \hat{\beta}_6\ln PA1564 \\
+ \hat{\beta}_7\ln PAOVER64 + \epsilon_1
$$

4.1.6. Multicollinearity

In a multiple regression model, imperfect multicollinearity is a phenomenon where two or more predictor variables are highly correlated – but not perfectly correlated (Stock and Watson, 2012). If so, the coefficients on at least one individual regressor would be imprecisely predicted (Hair et al., 2010; Adkins and Hill, 2008). However, the overall fit of the equation will be largely unaffected by this problem. This dissertation has significantly diminished the impact of the multicollinearity by transforming the data and expressing in logarithms (see section 4.1.3).

4.1.7. Heteroscedasticity

Although heteroscedasticity does not lead to biased parameter estimates, it violates the OLS assumption that errors are both independently and identically distributed. As a result, the estimated SE will be incorrect and confidence intervals with the desired confidence level cannot be produced.
(Hayes and Cai, 2007). In order to deal with this problem, this dissertation has made advantages of Stata.13 to estimate heteroscedasticity-robust standard errors.

4.1.8. Shapiro-Wilk W test

The test will be performed to validate the normality of the data (Ghasemi and Zahediasl, 2012). It is a fundamental condition in the further statistical analysis, because their validity depends on the normality of the data. The null-hypothesis of this test is that the data used in this dissertation follow a normal distribution. If the p-value is greater than the chosen level, which is 0.05 in this case, then the null hypothesis cannot be rejected and there is evidence that the data tested are from a normally distributed population (Ghasemi and Zahediasl, 2012). In Table 2, the p-value of 0.0556 is significant at p< 0.05, which is sufficient to establish normality of the data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>W</th>
<th>V</th>
<th>z</th>
<th>Prob&gt;z (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>33</td>
<td>0.9378</td>
<td>2.124</td>
<td>1.567</td>
<td>0.05856</td>
</tr>
</tbody>
</table>

This result is also confirmed by the graphical assessment of normality (see Figure 7). The figure plots the quantiles of residuals against the quantiles of a normal distribution. It can be seen that residuals from the database are reasonably normally distributed.

Figure 7: Quantile Plot
4.1.9. Pearson Product Moment Correlation Test

The previous efforts have provided the sufficient conditions in order to produce the Person’s correlation test. This test not only can examine whether there is an association between each explanatory variables and CAHS, but also the strength of this relationship that exists between them. From Table 4, all explanatory variables (left-side) are significant at the 1% level of significance, which implies these variables play an important role in explaining CAHS, as said in the literature. Thus, it reflects the variables selected in this dissertation are relatively satisfied.

### Table 5: Pearson’s Correlation Test

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation Coefficient</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(CAHS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFR</td>
<td>-0.9171***</td>
<td></td>
</tr>
<tr>
<td>APG</td>
<td>-0.8962***</td>
<td></td>
</tr>
<tr>
<td>POMP</td>
<td>-0.7711***</td>
<td></td>
</tr>
<tr>
<td>ln(HETE)</td>
<td>0.9269***</td>
<td></td>
</tr>
<tr>
<td>ln(PA014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(PA1564)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(PAOVER64)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source(s): *** denotes statistically significant at 1% level of significance.

4.1.10. Inferential Test

The inferential statistics are useful in empirical evaluation. F-statistic and t-statistic, as the main hypothesis tests, will be employed in the further empirical discussion. Additionally, t-tests are performed to test for a single coefficient and F-tests aim to test joint hypotheses about regression coefficients. First of all, the null and alternative hypotheses need to be stated clearly before they can be tested. The two-sided alternative hypothesis used in this dissertation can be expressed as follows:

**Equation 2: Null and Alternative Hypotheses**

\[ H_0 : \beta = 0 \]
\[ H_1 : \beta \neq 0 \]

where \( H_0 \) and \( H_1 \) are the null and alternative hypotheses, respectively.

Moreover, the p-value will then be computed to examine whether the null hypothesis is rejected at three different significance levels - 1%, 5%, and 10% (Stock and Watson, 2012). Notably, the p-value
should be used only as guidelines and be treated as the tentative results until explained and confirmed by the relevant studies (Stock and Watson, 2012).

4.1.11. Goodness of Fit

According to Stock and Watson (2012), the values of R-squared and adjusted R-squared are used to qualify how the model fits the data. The R-squared measures how much variation of the dependent variable can be explained by the explanatory variables. The R-squared value tends to be higher by adding additional predictors in the model and thereby cannot state whether a regression model is adequate. Instead, the adjusted R-squared would be more relevant indicator of the goodness of fit of the model, because it would increase when the new added independent variable has a correlation to the dependent variable (Stock and Watson, 2012). The value of root mean squared error also provides a measure of overall performance of an estimator.

4.2. Results

The final regression results in Tables 6 have omitted superfluous variables, diminished the severity of multicollinearity, and corrected for the problem of heteroscedasticity. The columns labeled (1) through (5) each report separate regressions, which have the same dependent variable, ln(CAHS). Additionally, each explanatory variable follows the estimated regression coefficients, with their robust standard errors below them in parentheses. The constant is the predicted value of dependent variable when all explanatory variables equal zero. The final four rows include the summary statistics for regressions. They are F-statistics, adjusted R-squared, root mean square error and the sample size.

Table 6: Regression Results

<table>
<thead>
<tr>
<th>Dependent Variable ln(CAHS)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory Variables</td>
<td>TFR</td>
<td>APG</td>
<td>POMP</td>
<td>ln(TEHE)</td>
<td></td>
</tr>
<tr>
<td>TFR</td>
<td>-3.684***</td>
<td>-2.283***</td>
<td>-2.179***</td>
<td>-2.623***</td>
<td>-0.356**</td>
</tr>
<tr>
<td></td>
<td>(0.293)</td>
<td>(0.552)</td>
<td>(0.851)</td>
<td>(0.262)</td>
<td>(0.195)</td>
</tr>
<tr>
<td>APG</td>
<td>-2.136***</td>
<td>-2.145***</td>
<td>3.762***</td>
<td>0.446**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.660)</td>
<td>(0.691)</td>
<td>(0.565)</td>
<td>(0.207)</td>
<td></td>
</tr>
<tr>
<td>ln(TEHE)</td>
<td>-0.193</td>
<td>-0.699**</td>
<td>-0.346***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.665)</td>
<td>(0.225)</td>
<td>(0.062)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.011***
4.3. Discussion

As can be seen from Table 6, the first estimated regression only takes into account the TFR as an independent variable. The null hypothesis that the coefficient of TFR is zero is rejected at the 1% significance level. It follows that a one-unit decrease of TFR can expect CAHS to increase by 36.84%.

However, when an additional independent variable APG is added in the regression (2), the coefficient on TFR appreciably rises from -0.3684 to -0.2283 while it remains statistically significant at the 1% significance level. Simultaneously, the adjusted R-squared slightly increases from 0.8360 to 0.8692. This seems enough to warrant including APG in the regression as a deterrent to omitted variable bias. The APG variable is significantly different from zero at the 1% significance level and

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(PA014)</td>
<td>2.659***</td>
<td>(0.644)</td>
</tr>
<tr>
<td>ln(PA1564)</td>
<td>2.875**</td>
<td>(1.000)</td>
</tr>
<tr>
<td>ln(PAOVER64)</td>
<td>7.172***</td>
<td>(0.661)</td>
</tr>
<tr>
<td>Constant</td>
<td>17.450***</td>
<td>(0.606)</td>
</tr>
</tbody>
</table>

Summary Statistics

F-Statistics testing all variables=0
0.000

Adjusted R-squared

<table>
<thead>
<tr>
<th></th>
<th>0.8360</th>
<th>0.8692</th>
<th>0.8650</th>
<th>0.9851</th>
<th>0.9993</th>
</tr>
</thead>
</table>

Root Mean Square Error

<table>
<thead>
<tr>
<th></th>
<th>0.85556</th>
<th>0.76406</th>
<th>0.7763</th>
<th>0.25795</th>
<th>0.05542</th>
</tr>
</thead>
</table>

Number of Observations

<table>
<thead>
<tr>
<th></th>
<th>33</th>
<th>33</th>
<th>33</th>
<th>33</th>
<th>33</th>
</tr>
</thead>
</table>

Notes: * denotes statistically significant at 10% level of significance (p<0.1)
** denotes statistically significant at 5% level of significance (p<0.05)
*** denotes statistically significant at 1% level of significance (p<0.01)
Robust standard errors are in parentheses.
negatively associated with CAHS. A 213.6 per cent growth in CAHS occurs with each 1 per cent decline in APG, holding the TFR variable constant. However, the sign of the coefficient on APG has switched from positive in the regressions (2) and (3) to negative in the last two regressions. Thus, it can be argued that regressions (2) and (3) could suffer from the omitted variable bias, implying both education and age structure factors could be the important determinants of CAHS. According to the last two regressions, APG is significantly and positively associated with CAHS. This outcome can be explained by the arguments proposed by Ito and Rose (2010) and Cook (2006). In China’s context, the lower population growth rates induced by the OCP have caused losses in the working-age population, who are potential savers in an economy (Choukhmane et al., 2014; Zhu et al., 2014; Banerjee et al., 2014; 2010; Wang, 2009). As a result, decreased APG can contribute negatively to CAHS.

In addition to other variables in regression (2), POMP is added in regression (3). Comparing those two regressions, the coefficients on both TFR and APG are significant at the 1% significance level, but that on POMP is not statistically significant. Nevertheless, it becomes significant at the 5% level of significance, after adding ln(TEHE) into the regression (4). Apart from that, the adjusted R-squared increases from 0.8650 in regression (3) to 0.9851 in regression (4). Moreover, the coefficient on ln(TEHE) is zero is rejected at 1% level and positively associated with CAHS. These findings together indicate that the incorporation of ln(TEHE) in regression (4) can be able to provide a better model to explain CAHS.

On the one hand, it is observed that the statistical significance of POMP is altered and it becomes negatively related to CAHS. Although this contrasts with the results suggested by Li et al. (2011) and Wei and Zhang (2008), it agrees with the arguments proposed by Griskevicius et al. (2012), Weir et al. (2011), Kvarnemo and Anhesjo (1996), as well as Taylor and Bulmer (1980). They argued that aggregate household savings could decline in response to the male-biased sex ratio. With higher competitiveness in Chinese marriage market, men tend to enhance their attractiveness by increasing their conspicuous consumption during their courtship. Consequently, CAHS would decrease.

On the other hand, based on the results in regression (4), an increase in the number of people enrolled in the higher education by 1 percentage point is associated with an increase in CAHS by 1.093 percentage points, other variables being equal. The outcome is in accordance with the following interpretations by Choukhmane et al. (2014), Zhu et al. (2014), Rosenzweig and Zhang (2009), Angrist et al. (2005), Ginther and Pollak (2003), Behrman et al. (1994) and Becker and Lewis
(1973). They maintain that the fertility control can lead to a trade-off between quantity and quality of children and this hence could contribute higher CAHS. Parents are expected to receive financial support from their children for their retirement consumption. Thus, they would be encouraged to allocate more savings to invest in children’s education, especially given the fewer dependent children under the OCP (Schultz, 2004). This would also help to account for the negative association between TFR and CAHS, as expected.

The final regression further exposes the effects of demographic transition by attempting to account for the impacts of three age groups on CAHS. There are four arguments can be made from the regression (5). Firstly, the coefficient on ln(PA014) is statistically different from zero at the significance level of 1%. The positive coefficient indicates that a percentage point decrease in the population aged between 0 and 14 would lead to a decrease in CAHS by approximately 2.659 percentage points, after controlling for the other variables in the model. This result contradicts my hypothesis that the sign of the coefficient on PA014 would be negative, but it reinforces the hypotheses by Zhu (2011), Kelley and Schmidt (1995) and Weil (1994). The decline in the proportion of people of non-working age of 0-14 has been brought about a rapid decline in the TFR under the OCP. Hence, there will be a fall in the number of young population entering into the labour market, which then could reduce CAHS.

Secondly, the estimated parameter of ln(PA1564) is statistically different from zero at the 1% level of significance and also has the expected sign. The coefficient on the parameter of ln(PA1564) means that the effect of 1 per cent increase in number of people aged 16-54 is expected to have 2.875 per cent increase in CAHS, controlling for the other variables in the model. The results are in line with those obtained by Choukhmane et al. (2014), Zhu et al. (2014), Banerjee et al. (2010), Wang (2009) and Ando and Modigliani (1963). These studies suggest that the increased share of middle-aged working population, which has been classified as potential savers in the society, has the potential to stimulate CAHS. This manifests in three different ways. First of all, this age group would accumulate more capital to support and educate their children in order to obtain higher returns from their children when they retired. The second point is that few children to support them in retirement, which accordingly stimulates them to save more to secure their retirement consumption. As a final point, they are responsible for their old parents with financial transfers and this also motives them to save more (Choukhmane et al., 2014; Zhu et al., 2014; Banerjee et al., 2010; Curtis, 2011; Wang, 2009; Bloom and Williamson, 1998; Ando and Modigliani, 1963).
Thirdly, the estimated coefficient on ln(PAOVER64) is positive and significant at the 1% significance level, which is inconsistent with my expectations. The result demonstrates that if the retired population increased by 1 per cent, CAHS would increase by 7.172 percentage points accordingly, holding constant the other independent variables. However, this agrees with the investigations by Ito and Rose (2010), Modigliani and Cao (2004) and Kraay (2000). It has been interpreted from two different perspectives. According to Ito and Rose (2010), the population aged 65 and over in China strongly depends on public pensions and family financial transfer from their children, who are middle aged; this can positively contribute to CAHS. Alternatively, due to the OCP, the elderly have fewer children can rely on. They thus tend to accumulate excess savings for themselves (Zhu, 2011). Overall, it seems that these demographic changes can be powerful drivers of CAHS.

Fourthly, in order to more fully understand how the aforementioned four key factors (TFR, APG, gender ratio, demographic structure) affect CAHS, it is useful to compute an F-test to assess the significance of these four factors through testing coefficients of all dependent variables equal to zero versus at least one of them differs from zero. This is also why ln(TEHE) was dropped in the final regression. The regression (5) reveals that the full set of variables is jointly significant at the 1% level in explaining CAHS. Apart from that, the final regression yields a relatively higher adjusted R-squared with 0.9993, compared to 0.8360 in the first regression. These together indicate that the regression has greater explanatory power in explaining CAHS.

What is more, the estimated effects of TFR on CHAS change substantially from first to the final regression, and it always remain a negative sign and statistically significant at the 1% level of significance (see Table 6). This indicates that TFR is a potent predictor of CAHS. Moreover, its negative sign is in line with my hypothesis, which is also supported by many other studies (Choukhmane et al., 2014; Zhu et al., 2014; Rosenzweig and Zhang, 2009; Li et al., 2008; Angrist et al., 2005). In the regression (5), a one-unit decrease of TFR is estimated to raise CAHS by 35.6 per cent.

Overall, the aforementioned statistical evidence can be used to conclude that the OCP (indicated by the four factors) plays a significant role in contributing CAHS.

4.4. Limitations

Although every attempt has been made to ensure that the above regression models fit the purpose of this dissertation. The major limitations of the models were mainly caused by the problems of data used in this dissertation. There were variables that could not be included in the model due to the
lack of data during the period of 1980-2012. For example, it would have been useful to analyse household aggregate savings in urban and rural areas (Ge et al., 2012; Qian, 1998), as well as in different regions (Choukhanine et al., 2014). Subsequently, the regressions used in this study may suffer from omitted variable bias. The corresponding results might be biased (Stock and Watson, 2012) and the impact of the OCP on CAHS would thus be underestimated. This also has the potential to produce the high values of R-squared and Adjusted R-squared.

5. Conclusion

This empirical dissertation has investigated the potential relationship between the One-Child policy (OCP) (it was applied only to the Han Chinese) and China’s aggregate household savings (CAHS). The aim of this dissertation has derived from the fact that a decrease in the TFR as a result of the OCP is accompanied by a dramatic growing trend in CAHS from 1980 to 2012. The dissertation has, firstly, reviewed the theory and relevant literature in order to assist in the formulation of methodology and hypotheses. Specifically, the existing literature has helped identify four factors as the proxies for the OCP: the TFR, the annual population growth rate, the age structure, as well as the gender ratio. In addition, the theory of Life-Cycle Saving led by Modigliani (1986) has been taken as the theoretical foundation.

Accordingly, a multiple regression analysis has been constructed based on the cross-sectional annual data from the Chinese official database and the World Bank between 1980 and 2012. I have employed a set of methods with the purpose of generating more valid empirical results; particularly the Shapiro-Wilk W test, Person’s correlation test and Inferential test. The final empirical results indicate that not only the individual variable is statistically significant at the 1% level of significance; the set of variables is also jointly significant in estimating CAHS; TFR, the population aged 15-64, 65 and over are all statistically significant in explaining changes in CAHS. Moreover, according to the adjusted R-squared value, the model I created seems to fit the data well. These together have led to the conclusion that the OCP has had detectable effects on CAHS. The estimated signs of the coefficients on the youth and the elderly are inconsistent with the theory of LCS and my hypotheses. However, this result has been supported by many other studies. Due to the limitation in data, further improvements of the study are necessary to have a more satisfactory result, such as an analysis of classifying CAHS by different regions. Although this study has limitations, it is among the first to measure the impact of the OCP on CAHS. It has also overcome two main defects existing in
previous literature which often had a small sample size and isolated the impact of each factor on CAHS.
Bibliography


Appendix

Appendix A

Table A1: Data on aggregate household savings, the percentage of male population and the total enrolment in the higher education in China (1980-2012)

<table>
<thead>
<tr>
<th>Year</th>
<th>China's aggregate household savings (100 Million Chinese Yuan)</th>
<th>The percentage of male population (Percentage)</th>
<th>The total enrolment in the higher education (10,000 persons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>395.8</td>
<td>52.06475</td>
<td>114.4</td>
</tr>
<tr>
<td>1981</td>
<td>523.4</td>
<td>52.19492</td>
<td>127.9</td>
</tr>
<tr>
<td>1982</td>
<td>675.4</td>
<td>52.31433</td>
<td>115.4</td>
</tr>
<tr>
<td>1983</td>
<td>892.9</td>
<td>52.28717</td>
<td>120.7</td>
</tr>
<tr>
<td>1984</td>
<td>1,214.70</td>
<td>52.27555</td>
<td>139.6</td>
</tr>
<tr>
<td>1985</td>
<td>1,622.60</td>
<td>52.44018</td>
<td>170.3</td>
</tr>
<tr>
<td>1986</td>
<td>2,237.80</td>
<td>52.50872</td>
<td>188</td>
</tr>
<tr>
<td>1987</td>
<td>3,083.40</td>
<td>52.35938</td>
<td>195.8725</td>
</tr>
<tr>
<td>1988</td>
<td>3,819.10</td>
<td>52.33394</td>
<td>206.5923</td>
</tr>
<tr>
<td>1989</td>
<td>5,184.50</td>
<td>52.32918</td>
<td>208.2111</td>
</tr>
<tr>
<td>1990</td>
<td>7,119.60</td>
<td>52.26434</td>
<td>206.2695</td>
</tr>
<tr>
<td>1991</td>
<td>9,244.90</td>
<td>52.01123</td>
<td>204.3662</td>
</tr>
<tr>
<td>1992</td>
<td>11,757.30</td>
<td>51.64</td>
<td>218.4376</td>
</tr>
<tr>
<td>1993</td>
<td>15,203.50</td>
<td>51.61004</td>
<td>253.6</td>
</tr>
<tr>
<td>1994</td>
<td>21,518.80</td>
<td>51.67697</td>
<td>279.9</td>
</tr>
<tr>
<td>1995</td>
<td>29,662.30</td>
<td>51.57113</td>
<td>290.6</td>
</tr>
<tr>
<td>1996</td>
<td>38,520.80</td>
<td>51.35361</td>
<td>302.1</td>
</tr>
<tr>
<td>1997</td>
<td>46,279.80</td>
<td>51.58225</td>
<td>317.4</td>
</tr>
<tr>
<td>1998</td>
<td>53,407.47</td>
<td>51.72051</td>
<td>340.9</td>
</tr>
<tr>
<td>1999</td>
<td>59,621.83</td>
<td>51.85274</td>
<td>408.5874</td>
</tr>
<tr>
<td>2000</td>
<td>64,332.38</td>
<td>52.02248</td>
<td>556.09</td>
</tr>
<tr>
<td>2001</td>
<td>73,762.43</td>
<td>51.81509</td>
<td>719.07</td>
</tr>
<tr>
<td>2002</td>
<td>86,910.65</td>
<td>51.8033</td>
<td>903.36</td>
</tr>
<tr>
<td>2003</td>
<td>103,617.65</td>
<td>51.8135</td>
<td>1108.6</td>
</tr>
<tr>
<td>2004</td>
<td>119,555.39</td>
<td>51.82818</td>
<td>1333.5</td>
</tr>
<tr>
<td>2005</td>
<td>141,050.99</td>
<td>51.83171</td>
<td>1561.777</td>
</tr>
<tr>
<td>Year</td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>2006</td>
<td>161,587.30</td>
<td>51.79724</td>
<td>1738.844</td>
</tr>
<tr>
<td>2007</td>
<td>172,534.19</td>
<td>51.768</td>
<td>1884.895</td>
</tr>
<tr>
<td>2008</td>
<td>217,885.35</td>
<td>51.73505</td>
<td>2021.025</td>
</tr>
<tr>
<td>2009</td>
<td>260,771.66</td>
<td>51.69124</td>
<td>2144.657</td>
</tr>
<tr>
<td>2010</td>
<td>303,302.49</td>
<td>51.51592</td>
<td>2231.793</td>
</tr>
<tr>
<td>2011</td>
<td>343,635.89</td>
<td>51.5083</td>
<td>2308.508</td>
</tr>
<tr>
<td>2012</td>
<td>399,551.00</td>
<td>51.50481</td>
<td>2391.316</td>
</tr>
</tbody>
</table>

**Sources:** NBSC, 2013a, p.1; 2013b, p.1; 2013c, p.1.
Table B1: Data on the total fertility rate, annual population growth, population aged 0 to 14, aged 15 to 64, aged 65 and over in China (1980-2012)

<table>
<thead>
<tr>
<th>Year</th>
<th>Total fertility rate (Births per woman)</th>
<th>Annual population growth (Percentage)</th>
<th>Population aged 0 to 14 (10,000 persons)</th>
<th>Population aged 15 to 64 (10,000 persons)</th>
<th>Population aged 65 and over (10,000 persons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>2.71</td>
<td>1.254221</td>
<td>34735.99</td>
<td>58406.94</td>
<td>4756.15799</td>
</tr>
<tr>
<td>1981</td>
<td>2.673</td>
<td>1.280952</td>
<td>34204.33</td>
<td>60024.56</td>
<td>4931.057753</td>
</tr>
<tr>
<td>1982</td>
<td>2.682</td>
<td>1.472675</td>
<td>33713.93</td>
<td>61797.03</td>
<td>5119.619005</td>
</tr>
<tr>
<td>1983</td>
<td>2.72</td>
<td>1.44495</td>
<td>33238.89</td>
<td>63551.65</td>
<td>5312.366206</td>
</tr>
<tr>
<td>1984</td>
<td>2.769</td>
<td>1.312069</td>
<td>32806.32</td>
<td>65165.89</td>
<td>5501.651272</td>
</tr>
<tr>
<td>1985</td>
<td>2.811</td>
<td>1.361699</td>
<td>32521.4</td>
<td>66716.38</td>
<td>5690.59672</td>
</tr>
<tr>
<td>1986</td>
<td>2.826</td>
<td>1.487399</td>
<td>32425.56</td>
<td>68243.12</td>
<td>5875.318131</td>
</tr>
<tr>
<td>1987</td>
<td>2.806</td>
<td>1.603605</td>
<td>32505.87</td>
<td>69751.32</td>
<td>6050.379102</td>
</tr>
<tr>
<td>1988</td>
<td>2.745</td>
<td>1.610071</td>
<td>32707.33</td>
<td>71178.52</td>
<td>6206.60739</td>
</tr>
<tr>
<td>1989</td>
<td>2.644</td>
<td>1.53317</td>
<td>32968.39</td>
<td>72485.87</td>
<td>6342.563691</td>
</tr>
<tr>
<td>1990</td>
<td>2.506</td>
<td>1.467303</td>
<td>33260.33</td>
<td>73700.83</td>
<td>6468.358236</td>
</tr>
<tr>
<td>1991</td>
<td>2.342</td>
<td>1.364434</td>
<td>33564.9</td>
<td>74795.61</td>
<td>6594.867701</td>
</tr>
<tr>
<td>1992</td>
<td>2.171</td>
<td>1.225536</td>
<td>33857.9</td>
<td>75752.17</td>
<td>6729.382808</td>
</tr>
<tr>
<td>1993</td>
<td>2.009</td>
<td>1.149619</td>
<td>34113.36</td>
<td>76661.55</td>
<td>6878.947501</td>
</tr>
<tr>
<td>1994</td>
<td>1.865</td>
<td>1.130261</td>
<td>34293</td>
<td>77624.81</td>
<td>7045.75308</td>
</tr>
<tr>
<td>1995</td>
<td>1.746</td>
<td>1.086509</td>
<td>34344.04</td>
<td>78668.05</td>
<td>7227.531686</td>
</tr>
<tr>
<td>1996</td>
<td>1.656</td>
<td>1.048142</td>
<td>34261.19</td>
<td>79801.09</td>
<td>7422.45152</td>
</tr>
<tr>
<td>1997</td>
<td>1.591</td>
<td>1.02345</td>
<td>34049.73</td>
<td>81032.17</td>
<td>7629.846199</td>
</tr>
<tr>
<td>1998</td>
<td>1.546</td>
<td>0.95955</td>
<td>33672.45</td>
<td>82352.46</td>
<td>7846.793264</td>
</tr>
<tr>
<td>1999</td>
<td>1.52</td>
<td>0.865851</td>
<td>33095.62</td>
<td>83759.6</td>
<td>8071.606163</td>
</tr>
<tr>
<td>2000</td>
<td>1.51</td>
<td>0.787957</td>
<td>32318.23</td>
<td>85271.96</td>
<td>8304.808897</td>
</tr>
<tr>
<td>2001</td>
<td>1.514</td>
<td>0.726381</td>
<td>31331.82</td>
<td>86913.83</td>
<td>8546.728028</td>
</tr>
<tr>
<td>2002</td>
<td>1.527</td>
<td>0.67</td>
<td>30169.39</td>
<td>88658.49</td>
<td>8796.284356</td>
</tr>
<tr>
<td>2003</td>
<td>1.546</td>
<td>0.622861</td>
<td>28931.01</td>
<td>90423.64</td>
<td>9055.666334</td>
</tr>
<tr>
<td>2004</td>
<td>1.566</td>
<td>0.593933</td>
<td>27752.12</td>
<td>92105.23</td>
<td>9324.903869</td>
</tr>
<tr>
<td>2005</td>
<td>1.585</td>
<td>0.588125</td>
<td>26736.35</td>
<td>93633.4</td>
<td>9598.135823</td>
</tr>
<tr>
<td>------</td>
<td>----------------------</td>
<td>------------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>2006</td>
<td>1.602</td>
<td>0.558374</td>
<td>25910.11</td>
<td>94953.42</td>
<td>9862.593102</td>
</tr>
<tr>
<td>2007</td>
<td>1.617</td>
<td>0.522272</td>
<td>25259.44</td>
<td>96065.69</td>
<td>10110.92549</td>
</tr>
<tr>
<td>2008</td>
<td>1.63</td>
<td>0.512387</td>
<td>24784.98</td>
<td>96993.56</td>
<td>10344.96615</td>
</tr>
<tr>
<td>2009</td>
<td>1.642</td>
<td>0.497381</td>
<td>24462.21</td>
<td>97744.56</td>
<td>10569.55</td>
</tr>
<tr>
<td>2010</td>
<td>1.65</td>
<td>0.48296</td>
<td>24269.46</td>
<td>98331.19</td>
<td>10792.24269</td>
</tr>
<tr>
<td>2011</td>
<td>1.657</td>
<td>0.47915</td>
<td>24212.63</td>
<td>98763.89</td>
<td>11024.79095</td>
</tr>
<tr>
<td>2012</td>
<td>1.663</td>
<td>0.487231</td>
<td>24291.4</td>
<td>99054.57</td>
<td>11278.31665</td>
</tr>
</tbody>
</table>

**Sources:** The World Bank, 2015, p.1; 2014c, p.1; 2014d, p.1; 2014e, p.1; 2013, p.1.
How Has Inequality Been a Cause of Violence in Post-dictatorship Brazil?

Samson Hart

This thesis explores the relationship between violence and inequality, beginning with an assessment of the historical context of violence and inequality in Brazil, and a review of the literature on these topics. The thesis goes on to qualitatively assess relevant theoretical frameworks, namely from Amartya Sen on inequality and Johan Galtung on violence, to develop a theoretical understanding of both the concepts of inequality and violence, in order to provide a more focused conceptualisation and a basis for their measurement. Finally, this thesis sets out a quantitative case study analysis of the relationship between violence and inequality in Brazil from 1985-2012, referring to different sources of data. The paper concludes that, given the limitations of the study, the data does not necessarily prove a causal relationship between inequality and violence. It does, however, provide enough evidence to call for the use of redistributive economic policy as a long-term preventative measure of violence.
1. Introduction

1.1 Historical Context

In order to grasp a true understanding of the current problems within Brazilian society, it is important to provide a historical context and explain the political, economic, and societal transition of Brazil in the years preceding the country’s democratisation. Brazilian violence is a phenomenon that must be understood from its roots and related back to the historical inequalities that characterise Brazilian society (Richardson and Kirsten, 2005). This section will thus set the scene and explain why the conditions of the past are tantamount to the violence and inequality facing Brazil today.

Latin America has been associated with violence for a long time, and some argue that it has become ingrained into the social structures. Ayers (1998) even describes the existence of a ‘culture of violence’ within the region. In Brazil, academics have linked the birth of this culture to the violent nature of the Portuguese colonisation (Lockhart, 1983; Kurtz, 2009; Miller, 2006). However, unlike other Latin American countries, Brazil managed to achieve some stability in the mid-19th century, gaining power and wealth through the utilisation of exports in a period of liberalised world trade, leading to a limited democratisation (Smith, 2002).

Brazil went on to pursue an ambitious industrialisation and development policy, which eventually ran into difficulty when the militarily infamously staged a coup d’état in 1964 (Mendes, 2015). This caused widespread institutionalised violence and sparked a long period of state terror in Brazil, reproducing the aforementioned ‘culture of violence’ (Imbusch et al., 2011). This period was equally important for the intensification of economic inequality within the region. The type of economic policy and political strategy adopted by the dictatorship created a privileged class and neglected the basic education of the masses (Mendes, 2015).

In fact, by the mid-1980s, Brazil was considered a successful case in terms of economic growth, but “a failure in terms of income distribution, social assistance, and poverty alleviation” (Mendes, 2015, p.6). As shown in Figure 1, Brazil’s per capita GDP started to rise by 1985, but it had a Gini index of almost 60 (World Bank, 2012a). Further to this, the income share held by the wealthiest 20% was 60% of the total, whilst the poorest 20% only received 2.88% of the wealth (see Figure 2). Additionally, at the time, Brazil had a severely uneducated population, an extremely low life expectancy (64 years), and over 30% of the population was living on less than two dollars a day (World Bank, 2015a).
During the decades of dictatorship, a shift in political and economic policy towards neoliberalism was accompanied by a change in the form of violence. In the past, the violence had been mainly politically motivated, i.e. used to “obtain or maintain political power” (Imbusch et al., 2011, p.89). However, in the 1970s and early 1980s, the political violence of the past decreased, with a movement towards the social, criminal and everyday violence that we are confronted with today (Brysk, 2003).
The transition from military dictatorship to democracy in 1985 brought about further changes to the country’s character but, frustrating any expectations, violence in the form of homicide and crime rates have escalated dramatically (see Figure 3). It is estimated that approximately 50,000 people are murdered every year in Brazil (DATASUS, 2012a), which represents only a small part of the much broader scenario of violence that has emerged over the past few decades (Rose, 2005).

Figure 3: Homicide trends in Brazil, 1985-2012

Democratisation has also done very little to tackle the severe problems of inequality in Brazil, which remains a country with one of the largest disparities between the rich and poor in the world (see Figures 1 and 2). While democracy has led to a redistribution of freedom, power and choice, it does not necessarily follow that democracy will bring a redistribution of wealth and income. This is especially true in Brazil, where the economic elite have a grip on governmental decision, and where great conflict exists between high-income groups that aim to maintain privileges obtained throughout the dictatorship, and low-income groups, who demand a redistribution of income and poverty relief (Mendes, 2015). Whilst the turn of the millennium saw government attempts to redistribute resources to the poor (see Figures 1 and 2), inequality is still rife, the poor are still extremely poor, and it is still the underprivileged groups in Brazil that are the main victims and perpetrators of violence (Ramos and Musumeci, 2005).

1.2 Outline

This dissertation will investigate how these endemic inequalities have been a cause of the violence in Brazil. Section 2 will critically review the relevant literature on violence and how it relates to inequality.
Following on from this, Section 3 will develop the theoretical perspectives on violence and inequality that are essential to their understanding, both as concepts and in how they relate to each other. Section 4 will outline the case study methodology, its aims, and the motivation and reasoning behind this choice of methodology. The results of the case study will be outlined in Section 5, drawing upon the relationship between income inequality and different measures of violence in different contexts. Finally, Section 6 will conclude the dissertation, tying together the theoretical and empirical aspects in order to outline relevant policy recommendations and highlight the appropriate areas that must be developed for further research.

2. Literature Review

2.1 Introduction

The aim of this Section is to provide a critique and overview of the academic literature that addresses both inequality and violence in Brazil. This Section will thus present the situation of violence in Brazil, and the changes of the academic understanding of violence and its causes. The concentration will be on the violence literature, for its transformation over the past few decades says a lot about the changes in how we have come to understand violence and its relationship with inequality.

There are two factors that must be accounted for in understanding the current state of violence research. First, it is not possible to interpret violence in the same way for different forms. The Israel-Palestine conflict, World War II, serial murders or the urban violence of Brazil are all forms of violence existing in separate contexts (Arendt, 1970), and thus the literature on violence research in Brazil points to an insurmountable number of factors that cause violence in the region. And second, as “no single factor is able to adequately explain the high levels of violence” (Imbusch et al., 2011, p.119), there is significant debate over which factors are the main causes of violence. This section will thus attempt to outline the academic literature in the context of some different disciplines and in relation to different aspects of inequality, before providing a framework that combines these viewpoints into a sociological model.

2.2 Violence as Interdisciplinary

Traditionally, violence was recognised as an issue for the criminologist and treated as a kind of deviant behaviour that deserves punishment, as opposed to a reaction to circumstance or even human nature (Arias, 2006). Some economists agree with this view, and there has been a vast amount of research on the probability of being caught as a large factor in determining levels of violence (Becker, 1968; Stigler, 1970; Ehrlich, 1973). However, there is much evidence that punishment and law do very little
to deter perpetrators of violence. In fact, some argue that the police force in Brazil act as an additional security risk and increase the fear of victimisation (Brinks, 2007; Tulchin and Ruthenburg, 2008). Murray et al. (2013) found that nearly 2,000 people were killed by the state police or military in Brazil in 2010 (1.6 per 100,000 of the population), a homicide rate higher than the total in the UK and France. It is clear that strategies used by state police are highly excessive and ineffective. Thus, police reform is an urgent matter, and whilst violence today is largely a criminal phenomenon, understanding violence from a purely criminal perspective is inadequate (Greene and Pranis, 2007; Uildirks, 2009).

Political scientists and sociologists have attempted to analyse the issue of violence within their own research frameworks. They look at the social background of perpetrators, analysing the internal logic underlying particular acts of violence and the political context of those violent acts (Imbusch et al., 2011). The literature on political violence does show links between the sociological behaviour of individuals and their willingness to commit acts of political violence. Some found positive relationships between political violence and inequality in the form of relative wealth, and the tendency of individuals to equate their grievances with such gaps (Nagel, 1974). Others found that violence in the political context was far more related to overall social well-being (Sigelman and Simpson, 1977). Political violence does still exist in Brazil, and is rooted in an exclusionary and non-egalitarian agricultural socioeconomic system. It occurs in rural areas, where structural conflict has existed for many years and has recently brought about new violent movements that are protesting and fighting for land reform and redistribution (Kay, 2000). However, the nature of violence in Brazil has undergone a transformation since the formation of democracy, and whilst violence continues to exist within the context of politics, most forms today are in a social or cultural context (Brysk, 2003).

Due to the large social and economic costs associated with violence and crime, violence is also considered to be a public health or macroeconomic issue. For macroeconomists, the costs associated with violence are detrimental to the economic development of societies. For the public health specialist, violence is considered a pandemic for its socially corrosive characteristics. The literature is in agreement that the high societal costs of violence “far exceed individual consequences and pose considerable costs on the country’s economy and its development” (Briceño-León et al., 2008, p.754). Violence erodes labour through limiting access to jobs, human capital in limiting access to education and health and social capital through reducing trust (Moser and Shrader, 1999; Heinmann and Dorte, 2006; Wilkinson, 2004). Morrison, Buvinic and Shifter (2003) present a typology for the socioeconomic costs of violence in Brazil, and found that the total cost of violence to society is 10.5% of GDP. Whilst this is an extremely high figure compared to the rest of the literature, it is still unanimously accepted
that violence has a devastating effect on Brazil’s overall GDP (Briceño-León et al., 2008; Murray et al., 2013; Mendes, 2015; Moser and Shrader, 1999). Whilst this literature highlights the need for policy measures to deter violence, it does not entail much about the true causes of Brazilian violence.

2.3 Violence and Inequalities

Research on inequality has come to the forefront in both developed and developing countries and it is no longer mere speculation that economic inequality has a relationship with severe social destruction. Literature that expresses a correlation between inequality and violence is vast within the region of Latin America. There are many economic studies that suggest income inequality, as opposed to poverty, is a large determinant of violence (Fajnzylber et al., 2002; Briceño-León et al., 2008; Portes and Roberts, 2005; Menjivar; 2008). Wilkinson and Pickett (2009) note that countries facing the strongest social inequalities are most likely to have the greatest problems with violence, and severe inequality most likely leads to violence. Newman (1999) backed these claims, proposing that a 1% rise in a country’s Gini coefficient is associated with a similar increase in its homicide rate.

Nevertheless, it is important to investigate the intersectional nature of the Brazilian violence phenomenon, in order to truly understand the relationship between inequality and violence in the region. In Brazil, different groups experience violence differently. For low income groups, homicides and physical assaults are far more common than for middle- and high-income groups, who are more likely to be affected by property crime (Imbusch et al., 2011). Brazil’s violence disproportionately affects young people, often occurring in the context of gangs, as a way of life to which there are few alternatives (Rodgers, 1999; Jones and Rodgers, 2009; Barker, 2005). In 2009 the homicide rate in Brazil was 62.5 per 100,000 people aged 20 to 29, a figure much higher than in any other age group (Murray et al., 2013). Within this form of gang violence, it is also males that are most effected, and the homicide rate is usually ten times higher than that of women’s (Krug et al., 2002; PAHO, 2006; Murray et al., 2013). Finally, rises in Brazilian homicides are also related to race, as Goldstein (2003) found that victims of violence have a tendency to be black. Hence, the literature expresses a need to understand that there inequalities in how violence affects different people.

The literature also articulates that, in the urban setting, violence is far more common, and urbanisation and rapid migration to the cities has had a large effect on the recent upsurge in violence (PAHO, 1996). These changes in patterns of urban criminality, since the mid-1970s, have led to many different interpretations. However, it is greatly accepted that urbanisation is interlinked with the creation of severe social inequality, leading to exclusion, frustration and violence (Concha-Eastman,
2001; Briceño-León, 2005). In essence, the literature shows that if you are a young, black male living in an urban, low-income neighbourhood in Brazil, it is likely that you will be a victim or perpetrator of violence.

2.4 A Framework to Understand Latin American Violence

The literature has shown that attempting to understand violence as something measurable and treatable with simple remedies is too simplified. In order to understand and treat violence we must recognise the social ecology that exists within societies and we must be able to categorise different risk factors in order to expose the root causes of violence in Brazil. The literature expresses the importance of understanding violence as a dynamic process involving many actors with many different causes and consequences. This is captured well in a model which classifies risk factors and, whilst many of these models are suggested in the literature (WHO, 2002; Moser and Shrader, 1999; Concha-Eastman, 2001; Bandura, 1973; Reiss and Roth, 1993), they fail to come to terms with the multifaceted and subjective nature of Latin American violence.

In order to truly understand the Latin American violence phenomenon, it is important to use a model that is created purely to do so. Briceño-León’s (2005) sociological model is not a universal explanation for violence, but rather a model that has been developed to provide an explanation for current violence in Latin America. The model is an interdisciplinary approach to the violence phenomenon that recognises the contribution of many different explanatory proposals, sharing aspects with all of them (Briceño-León, 2005). Briceño-León (2005) illuminates a framework that identifies three separate levels on which Latin American violence operates. As shown in Figure 4, each of these levels covers the next, encompasses it, and contributes to its conception.

The first level, factors that originate violence, is structural, and it refers to the processes that are created and sustained over a long period of time. These factors do not necessarily determine what occurs, but they are the social and cultural characteristics that create the basis for violent behaviour to persist (Briceño-León, 2005, 2006; Briceño-León et al., 2008; IDB, 1999). This level includes factors such as a lack of employment opportunity for young people, or a heightening in expectations without an ability to meet them, thus, they are characterised by extremely pervasive inequalities. Secondly, factors that foment violence refer to the meso-social influences that have more immediate effects on behaviour. This includes aspects of Latin American society, such as an increase in urban density, that promote violent action within a framework already characterised by the factors from which violence originates. Finally, there are factors that facilitate violence. These are not necessarily direct causes but
they influence the occurrence or lethality of violence. For example, the social norms that promote alcohol use or the carrying of firearms will have an impact on the quantity of violence occurring, and its lethality (Briceño-León, 2005; Briceño-León et al., 2008; Murray et al., 2013). Thus, within this model, violence is understood as the product of multiple and overlapping levels of factors, whereby the root causes are inequalities in society.

**Figure 4: The Causes of Latin American Violence: A Sociological Framework**

![Diagram](image)

Source(s): Briceño-León, 2005, p. 1633

To summarise, this section has reviewed the literature on violence, revealing the interdisciplinary nature of the academic understanding of violence, and the limitations of each individual discipline. It has shown the significance of understanding violence using a framework that appreciates the complex nature of the violence phenomenon. This dissertation will attempt to add to the existing body of literature by isolating inequality and assessing its relationship with Brazilian violence in an attempt to decipher the extent to which violence in Brazil is dependent on the inequality of Brazil.

**3. Theoretical Perspectives**

This section will develop the theoretical understanding of both the concept of violence and of inequality in order to provide a more focused conceptualisation and a basis for their measurement.

**3.1 On Violence**

The literature review illuminated the interdisciplinary nature of violence, helping to show the difficulties in its conceptualisation and establishing a need to understand the complex relationship
between the different causes of violence. This section will attempt to theorise violence as a concept, because the general understanding of violence is empirically grounded, rather than theoretically conceived.

Georges Sorel’s (Sorel, 1906, p.60) admission that “the problems of violence still remain very obscure” continues to have integrity today, and it is widely accepted that violence is a slippery concept that is in need of careful deliberation. Nevertheless, Bourgois and Scheper-Hughes (2004) depict a number of attributes of violence that are integral to understanding its relationship with inequality. First, violence must be understood in relation to human beings as social creatures and thus it is cultures, social structures, ideas, and ideologies that shape all dimensions of violence (Bourgois and Scheper-Hughes, 2004). Secondly, many acts of violence can be misrecognised or sometimes not even viewed as violent at all. For the socially marginalised groups these are instances of disease, starvation and humiliation and, whilst tracing those responsible becomes problematic, the violent nature of these conditions cannot be undermined (Bourgois and Scheper-Hughes, 2004).

In fact, these forms of violence that often go unnoticed tend to occur more often and with more devastation than physical forms of violence. Johan Galtung coined the term structural violence to describe the violence that is not directly inflicted from one person to another. Rather, it is “built into the structure and shows up as unequal power and consequently as unequal life chances” (Galtung, 1969, p.171). Structural violence is the existence of a political economy of inequality under neoliberal capitalism that promotes social suffering (Menjivar, 2008). This political economy will cause poverty and mortality and, whilst it does not directly cause interpersonal violence, it conditions structures within which people resort to inflicting pain on one another (Bourgois, 2004; see Figure 5). For example, in a society where there is unequal access to resources, causing a lack of opportunity for the lower classes, violence is exercised even if there is no traceable actor (Galtung, 1969). For the people in the lower classes, the structures of poverty and inequality that they receive is the true violence, and their crime, delinquency and personal violence is merely a manifestation of these structures (Minayo, 1994).

It is important to understand that Galtung uses different terminology to express essentially the same thing. Galtung (1969) explains that structural violence is synonymous with social injustice and, principally, with inequality. Thus, when talking about the relationship between inequality and violence, we are really talking about the relationship between structural violence and personal violence, or the effects of social injustice on interpersonal violence (Galtung, 1969).
The writings of Hannah Arendt (1970) also recognise the invisible nature of these forms of violence through untraceable domination. She describes the existence of a bureaucracy in which nobody can be held responsible, and this issue is “amongst the most potent causes of the current world-wide rebellious unrest, its chaotic nature, and its dangerous tendency to get out of control and run amuck” (Arendt, 1970, p.39). The severe social unrest and eventual violence amounts from the fact that those who stand to lose in a situation of inequality have nobody to blame, married with the existence of a natural human “disinclination to have power exercised over themselves” (Mill, 1861, p.65). Throughout history, where the repressed and marginalised are confronted with outrageous conditions or events, resorting to violence is tempting because of its intrinsic imminence and speed (Arendt, 1970).

These theories help to explain why the daily expressions of violence in Brazil are “linked to the broader structures of inequality that promote interpersonal violence” (Menjivar, 2008, p.7). This dissertation
will attempt to provide empirically grounded evidence for this relationship, and add to the credibility of the theoretical perspectives. Any attempt to neglect structural violence, in a region with “profound social inequalities and important class and race cleavages, with strong discrimination and social exclusion, with extreme wealth and extreme poverty” (Imbusch et al., 2011, p.89) would be to misunderstand the violence phenomenon entirely. Thus, it is of great importance to acknowledge the relationship between physical or interpersonal violence and structural violence, for they are in many ways interwoven.

However, when coming to measure violence against inequality, we refer to its physical and traceable forms, i.e.:

“The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community that either results in or has a high likelihood of resulting in injury, death, psychological harm, mal-development, or deprivation” (Krug et al., 2002, p.5).

This definition is suitable, as it distinguishes between the measurable and the immeasurable. In accepting that “the general formula behind structural violence is inequality” (Galtung, 1969, p.175), finding a relationship between personal forms of violence, such as homicide, and inequalities, such as income inequality, will illuminate the true causes of violence in Brazil to be the structures of inequality in society.

3.2 On Inequality

However, the complexities of inequality must also be considered, for there are many different inequalities that exist in societies, and there is much debate about the appropriate choice of space in which inequality should be measured.

Amartya Sen is a principal writer on matters of equality and his writings have been influential to the understanding of inequality and how it should be measured. He explains that the early writings on the national accounts\(^1\) and economic prosperity strongly reflect the Aristotelian principle of a need for the assessment of the causal influences on the conditions of living (Sen, 1997). In fact, the foundations of economics were focussed on the quality of life, rather than income or wealth. Sen articulates this, stating that:

---

\(^1\) These are the first estimates of national income, devised by William Petty in 1665. See Petty, 1899.
“Whilst the national accounts...established the foundations of the modern concept of income, the focus of their attention was never confined to this concept. They were also very aware of the basic issue that the importance of income is instrumental and circumstantially contingent rather than intrinsic and categorical” (Sen, 1997, p.392-393).

Thus, a focus on inequality as an entirely income-based phenomenon is naïve, as income should always be seen as a means to some other end, rather than an end in itself. However, Sen also recognised the shortcomings of measures of inequality that encompass wider scope, and the challenges in finding an appropriate notion of inequality that is “both theoretically adequate and empirically usable” (Sen, 1997, p.385).

The use of mainstream welfare economics for inequality analysis is problematic, for it is concerned with a set of questions that tend to avoid judgements on distribution altogether. In fact, Pareto optimality was essentially designed in order to remove the need for distributional judgments (Sen, 1973). As a Pareto-improvement is defined by a change that makes no one worse off and someone better off, if the poor cannot be made any better without cutting into the wealth of the rich, the situation would be Pareto-optimal, despite the existence of disparity between the rich and the poor (Sen, 1973).

The utilitarian approach is equally frail. Since utilitarianism sees the sum of individual utilities as the ultimate measure of social welfare, rises in overall utilities (or economic growth) that only affect the extremely well-off will be seen as positive. This approach provides a somewhat limited account of individual well-being, for it is simply not enough to have a greater sum of utilities. Additionally, these approaches make no attempt to contemplate the freedom of individuals to pursue well-being, because a person’s utility is not representative of their capabilities to convert resources into functionings, i.e. the valuable activities and states that make up people’s well-being (Sen, 1987).

Thus, Sen calls for a more well-rounded understanding of inequality that goes beyond the mainstream measures, highlighting a need to consider the heterogeneity of human beings and the diversity of variables in terms of which equality can be judged. Human beings are incredibly diverse, both in their personal characteristics (e.g. age, sex) and external characteristics (e.g. income inheritance, natural and social environment), and these diversities play a crucial role in determining individual capabilities to convert resources into functionings (Sen, 1995). These diversities also have an effect on the outcomes of different ways of measuring equality. For example, equality of incomes can result in the
inequality of well-being, due to the existence of human diversities. Correspondingly, a libertarian approach may give priority to the equal granting of liberties, thus automatically rejecting equality in the distribution of income (Sen, 1995).

These trade-offs beg the question: ‘Equality of what?’ This is particularly relevant when taking into account the situation of Brazil, where an introduction of equality of liberties through democratisation has been met by widening inequalities elsewhere (Mendes, 2015). It is important to stress that the choice of space in which inequality should be measured will always have an effect on other spaces (Sen, 1997), and thus it is of great importance to recognise the motivation for research in choosing an inequality measurement (Sen, 1995).

As the appropriateness of a particular space is ultimately dependant on the underlying purpose of the evaluation of inequality, the choice of space and selection of particular inequality measures within that space must be made representative of that purpose (Sen, 1995). Notably, the argument for paying attention to functionings in assessing inequalities “must not be seen as an all-purpose” (Sen, 1995, p.89) option. The purpose of this dissertation is to establish the extent to which inequality is a cause of violence, and so it calls for the need to use a space in which there exists a quantifiable measure of inequality. Sen recognises this, stating that “practical economics, no less than politics, is that of the possible, and that issue remains even when the need for going beyond income inequality is well accepted” (Sen, 1997, p.390). Thus, as long as there is an appreciation and evaluation of the limited application of income inequality in assessing equality in other spaces, choosing income inequality as a means for evaluation is fair.

3.3 Violence and Inequality: The Inherent Link

If we can also accept income inequality as a form of structural violence, the conclusions of a relationship between income inequality and forms of personal or physical violence, as defined by the WHO (Krug et al., 2002), can add to the credibility of the theoretical literature on these concepts. As inequalities have a tendency to be experienced unilaterally (Galtung, 1969), and as different forms of violence have a tendency to produce and reproduce each other (Bourgois and Scheper-Hughes, 2004; Briceño-León, 2005), a concentration on the relationship between a particular form of inequality and a particular form of personal violence can still have relevance elsewhere. This would also serve as empirical backing for Briceño-León’s (2005) theoretical model of Latin American violence discussed in the previous Section, which asserts that inequalities are the true factors from which violence originates.
4. Methodology

This Section will aim to explain and justify the methodology used in this dissertation. First, the focus will be on the choice of a case study as the most appropriate methodology. Secondly, the choice of Brazil as a focus country will be addressed, as well as the reasoning behind choosing 1985-2012 as a time period. Finally, this section will outline the stages of the case study, their aims and the reasoning behind each stage.

4.1 Why a Case Study?

It is appropriate to use a case study when asking a how or why question that focuses on contemporary events over which the investigator has no control (Yin, 2009). In this case, we will be focusing on contemporary patterns of violence and inequality in Brazil, and asking how and why violence and inequality interlink. Case studies are used to motivate research through the questioning, clarification or illustration of existing theory (Siggelkow, 2007). This is appropriate, for the case study will be used to add an empirical element to the theoretical perspectives discussed in previous Sections. A case study also gives the author the opportunity to gain an in-depth understanding of a small number of cases or a specific case (Yin, 2009). Again, this is relevant for we will look at the particular effects of income inequality on specific forms of violence in Brazil.

Critics of case study methodology have argued, however, that case studies cannot provide reliable information about the broader class of phenomena in the area studied (Abercrombie, Hill and Turner, 1984). Flyvbjerg (2006) questions this claim, stating that “the force of example is underestimated” (p.228). Humans learn best from experience, thus, concrete, context-dependant knowledge is of great value to understanding. Case study knowledge is central to human learning (Christensen, 1987), so where proof is impossible, looking carefully at individual cases is the only way to learn something (Eysenck, 1976). Further criticism comes from the claim that case studies tend to have a subjective bias. However, a case study contains no greater bias than any other methodology, and, in actuality, case studies have “a greater bias towards falsification of preconceived notions than toward verification” (Flyvbjerg, 2006, p.237). The conclusions from case studies may well contradict the theory discussed, and thus it is unfair to say that there is a bias towards a particular theory or hypotheses.

4.2 Why Post-dictatorship Brazil?

In deciding on a country for the case study analysis, it is important to find one in which there exist a contemporary focus of inequality and violence. The recent riots in Brazil, a response to the hosting of
the World Cup in 2014, sparked global interest regarding the socially unjust nature of the event. It highlighted the “irreconcilable disparity between the scrummage of corporate interests that follows the World Cup and the concerns of those – citizens, fans, baffled and powerless observers – who form its backdrop” (Ronay, 2014). Additionally, despite a large amount of growth in recent decades, there have been persistent high levels of inequality in comparison with other countries. In 2011, Brazil’s GDP per capita grew to $11,320, the country’s highest ever GDP per capita, and yet its Gini Index remained at 53.09, a figure far from that of an egalitarian level (see Figure 1). Brazil also makes an interesting case when it comes to violence. Despite a lack of war and conflict in the region, violence occurs at a far greater rate than in the rest of the world, and even than in some countries in which there has been long-term conflict (Murray et al., 2013). The recorded homicide rate (see Figure 2) in 2008 was 29.6 per 100,000 people which, compared to the worldwide average of 7.9 (WHO, 2008), is astounding. Thus, there is great opportunity in Brazil to make strong findings regarding a relationship between violence and inequality.

The overall time period that will be considered in the case study is: 1985-2012. This time period represents the time period of democracy in Brazil, and thus, it will allow for an analysis into the lack of egalitarianism that democracy has brought about, contrary to the beliefs in political economic theory (Barro, 2000). The way in which violence is practiced in Brazil has also greatly changed following democratisation. As mentioned in previous Sections, the political violence of the past became the criminal and every-day violence that Brazilians experience today (Brysk, 2003), a violence that is more inherently linked to an unequal society. Additionally, the availability of data is strongest in this time period and it is essential to have reliable data in case study research (Yin, 2009).

4.3 Outline of Method and Aims

The case study analysis will include a number of stages. The first is to identify the existence or non-existence of a relationship between income inequality and the homicide rate over time. To show this, I will use data taken from the World Bank (2015a) for the Gini Index, and homicide figures from the Brazilian Ministry of Health (DATASUS, 2012a), which is “widely regarded as the most reliable information source on homicides in Brazil” (Murray et al., 2013, p.243). However, one of the largest criticisms of inequality and violence literature is the use of national level indicators where inequalities tend to occur on a more local basis, and where violence (especially in Brazil) is a localised, urban phenomenon (Østby, 2013; Cramer, 2001). The second stage will address this criticism by focusing on the income inequality and homicide figures for individual states.
There is also an issue with using homicide measures alone as an indicator for the levels of violence in Brazil, because the homicide rates only represent a small part of the much broader scenario of violence in the region (Rose, 2005). The next stage will thus, investigate other possible characterisations of violence, using victimisation surveys (Latinbarometro, 2012) and trends in hospitalisations for non-lethal violence (Brazilian Ministry of Health System). Assessing inequality against these figures will provide a much more realistic account of the relationship between inequality and violence in the region.

Another large criticism of the literature on inequality and its relation to violence is that other factors that are excluded are truly accountable for the changes in violence rates (Imbusch et al., 2011). Thus, the final stage of the case study will assess the effects of the external variables that are most commonly referred to as causes of Brazilian violence in the literature. These being: population growth, rates of urbanisation, overall GDP and levels of poverty.

5. Case Study: Income Inequality and Violence in Post-Dictatorship Brazil?

5.1 The Gini Index

The Gini index will be used as a measure of income inequality, for it is the most accurate and direct measure (Sen, 1973).

Figure 6: The Lorenz Curve

The Gini index can be measured using the Lorenz curve, where the horizontal axis represents the percentages of the population arranged from the poorest to richest, and the vertical axis represents the percentages of income enjoyed by each percentile of the population (Sen, 1973). This relationship
(see Figure 6) is representative of the Gini index, i.e. “the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line” (World Bank, 2015b, para.1). The data used has been scaled up, whereby a Gini index of 0 represents perfect equality, and a Gini index of 100 represents perfect inequality.

5.2 Homicide Trends and the Gini Index

In order to carry out this stage of the case study, data was collected from the World Bank (2015a) for the Gini coefficients. Measuring these figures each year from 1985 to 2012, alongside the corresponding homicide rates, that being the number of deaths “purposefully inflicted on a person by another person, expressed per 100,000 people” (UNDP, 2015, para.1), provides an indication of the relationship between the two variables as shown in Figure 7.

Figure 7: Gini Index vs. Homicide Rates, 1985-2012

![Graph showing the relationship between Gini Index and Homicide Rates](source)

Figure 7 represents a slight linear relationship between the rate of homicides and the Gini index. It shows that, contrary to the arguments of the theory discussed in previous sections, where income inequality is higher, the homicide rate is lower. This implies that there is in fact a negative relationship between inequality and violence. However, this relationship is relatively weak and it would be wrong to make these claims purely based on this one set of data. Additionally, this dissertation has expressed the limitations of these measures as representative of the real relationship between inequality and violence.
5.3 Homicide Rates and the Gini Index in Individual States

Using the national figures for income inequality and the homicide rates disregards the tendency for both inequality and violence to be localised phenomena. Figure 8 represents the uneven distribution of the rate of homicide in the states of Brazil, indicating that the national homicide rate is unrepresentative of this, and further investigation must be made into the relationship between the homicide rate and Gini index on a local level.

Figure 8: Homicide Rate in Brazilian States, 2009

![Homicide Rate in Brazilian States, 2009](source)

Figure 9: Gini Index vs Homicide for Individual States, 2009

![Gini Index vs Homicide for Individual States, 2009](source)

Figure 9 represents the relationship between each individual state’s Gini index and homicide rate in 2009. It shows that the states with the higher income inequality generally had a higher homicide rate. This is enlightening, for it indicates that, when accounting for the localised nature of the Brazilian
violence and inequality phenomena, the data shows a positive linear relationship between violence and inequality. For example, Alagoas had a relatively high Gini index (56.5) and a corresponding high homicide rate (59.3), whereas Santa Catarina had a relatively low Gini index (45.4) and a comparatively miniscule homicide rate (13.4). However, some states do not adhere to this trend. Acre, for example, had one of the highest levels of income inequality in Brazil (60.7), but a relatively low homicide rate (22.1). This shows that, whilst there is some empirical evidence here of a causal relationship between inequality and violence, more analysis must be made.

5.4 Alternative Measures of Violence

Homicide rates only represent a small part of the violence in Brazil, thus, this stage will investigate the broader scenario of violence in the region, by analysing the relationship between other measures of violence and the Gini index. First, survey data was taken from the Latinobarómetro, in which participants aged 16 and over were asked: “have you, or someone in your family, been assaulted, attacked, or been the victim of violence in the last 12 months?” (Latinobarómetro, 2012, cited in Murray et al., 2013, p.476). The results are given as a percentage of the participants who answered yes. These data, from the years 2001 to 2010, were then plotted against the corresponding Gini index in order to demonstrate whether or not there exists a correlation between the levels of income inequality and the likelihood of being a victim of violence.

Figure 10: Victimisations in past 12 months (%) vs Gini Index, 2001-2010

Source(s): World Bank, 2015a; Latinobarómetro, 2012
As shown in Figure 10, there is a positive correlation between the percentage of people who had been a victim of violence and the Gini index. Between 2001 and 2010, in the years when income inequality were greatest, so were the chances of victimisation. Generally, both victimisation and income inequality have been decreasing since the turn of the millennium. In 2010, victimisation rates were at the lowest point in the decade analysed (26.9%), whilst the income inequality was also the lowest in the time period (a Gini index of 53.34).

Second, data on the number of hospitalisations for violence that did not result in death, described as either resultant from assault or aggression, was extracted from the Brazilian Ministry of Health (DATASUS, 2012b). Whilst this is a relatively reliable source of data, it does not include data from private hospital admissions, which accounted for 30% of the total hospital admissions in 2008 (Murray et al., 2013). As shown in Figure 11, there is also a correlation between the number of hospitalisations for non-lethal violence and the Gini index in the years encompassing 1985 to 2010. Generally, in the years with higher income inequality, there are higher rates of hospitalisations for violence and, whilst these figures do not represent the full extent of hospitalisation rates, this finding adds to the credibility of an overall correlation between income inequality and violence.

Figure 11: Hospitalisations for Violence vs Gini Index, 1985-2010

Source(s): World Bank, 2015a; DATASUS, 2012b

Data before 2001 used a different sampling framework and was thus excluded.
5.5 External Causes of Violence in Brazil

The literature often refers to relationships between inequality and violence; however, there is also a tendency for the literature to view other factors as truly accountable for the levels of Brazilian violence. This stage will therefore explore the relationship between these external variables and rates of violence. The homicide rates will be used, as opposed to the other violence measures, based on their availability for the whole time period at question.

The population has grown from less than 140 million people in 1985 to almost 200 million in 2012 (World Bank, 2015a). At the same time, the percentage of this population living in an urban area has grown from 69.9% to 84.9%. When comparing the rate of urbanisation to the homicide rate, there is quite a strong correlation, suggesting that the higher the percentage of the population living in an urban area, the higher the homicide rate (see Figure 12). This correlation supports the evidence in the literature that suggests rapid migration to the cities has had a large effect on the upsurge in Brazilian violence (PAHO, 1998; Concha-Eastman, 2001). However, Briceño-León’s (2005) model argues that urbanisation merely foments violence, thus it is not a factor from which violence originates (see Figure 4).

Figure 12: Urban Population (% of total) vs Homicide Rate, 1985-2012

The percentage of the population living under $2 a day has dropped from 27.3% in 1985 to 6.8% in 2012 (World Bank, 2015a). In the same period, GDP per capita has also risen from US$1,637 to US$11,320. Figure 13 shows that, when comparing the rate of poverty and the rate of homicide, there is in fact a negative correlation suggesting that, where poverty is lower, violence rates are higher. Whilst this most probably does not indicate a causal relationship, it certainly does not support the
arguments in the literature that suggest the violence of Brazil is a result of absolute poverty levels (Imbusch et al., 2011). Figure 14 shows the relationship between GDP per capita and the homicide rate, suggesting that, contrary to the literature, rises in GDP per capita will not necessarily result in lower homicide rates.

**Figure 13: Poverty Rate (%) vs Homicide Rate, 1985-2012**

![Graph showing the relationship between poverty rate and homicide rate](image)

Source(s): World Bank, 2015a; DATASUS, 2012a

**Figure 14: GDP per capita vs Homicide Rate, 1985-2012**

![Graph showing the relationship between GDP per capita and homicide rate](image)

Source(s): World Bank, 2015a; DATASUS, 2012a

6. Conclusions

The aim of this dissertation has been to investigate and assess the relationship between inequality and violence, particularly in Brazil since its democratisation in 1985. This has been achieved through
the consideration of relevant literature and theoretical perspectives, and the use of a case study approach to identify the impact of changes in income inequality on different forms of violence in Brazil.

6.1 Summary of Findings

The case studies found that, when looking at the national figures on income inequality and their relation to national homicide rates, there is no indication of a positive correlation. However, when investigating the relationship between income inequality and the homicide rate for individual states, there is in fact a positive correlation. Further to this, there also appears to be a correlation between income inequality and other measures of violence, that being number of hospitalisations for non-lethal violence and chances of being a victim of violence. Finally, assessing the relationship between external factors and violence indicated that, whilst urbanisation rates and GDP per capita tended to increase alongside homicide levels, the rate of poverty was shown to be lower where violence is higher.

6.2 Connection to Theory

These findings contribute to the literature primarily though empirical backing of the existing theory. First, they provide empirical backing to Galtung’s (1969) theories regarding the true causes of violence and the relationship between structural and personal violence (see Figure 5). Essentially, Galtung’s theory states that the existence of a political economy of inequality in which violence is built into the structure of society is what truly accounts for social suffering (Galtung 1969). In Brazilian society, where there is severe income inequality, the lower classes have less opportunity and are conditioned into structures within which they resort to inflicting personal violence on one another. Thus, the key finding of this dissertation, i.e. that there is a relationship between forms of personal violence and income inequality, discerns the true causes of the Brazilian violence phenomenon to be the structures of inequality in society.

Additionally, whilst critics of case study methodology question its ability in aiding the understanding of the broader class of phenomena they underestimate the power of example and the truth that humans learn best from experience (Christensen, 1987). Sen (1995) was well aware that there must be an understanding of the tendency for inequalities to be experienced unilaterally, even in accepting the limited application of income inequality to the evaluation of equality elsewhere. Thus, the findings of a relationship between income inequality and violence can still be relevant in assessing the existence of a relationship between other inequalities and violence. Further to this, Briceño-León’s model (2005) describes a particular situation of Latin American violence in which inequality largely characterises the factors from which violence originates (see Figure 4). This model would argue that
the finding of a relationship between inequality and violence in Brazil is also relevant to the rest of Latin America.

6.3 Limitations and Recommendations for Further Research

There are, however, limitations to this research method and its results. First, a lack of rigor, in terms of sensitivity to details, may result in some irrelevant details being included, and this slightly lessens the credibility of a correlation. Secondly, the availability of data was relatively weak due to issues in collecting data from a Portuguese-speaking country. There are also several limitations, which highlight issues that could be developed in future research. There is a need for a better and more efficient measure of inequality than the Gini coefficient. A measure that incorporates not just income disparity, but other issues such as the distribution of power and the capabilities of individuals, would be far more representative of the full extent of inequality. Further to this, there is a need for more trustworthy and rigorous measures of overall violence, especially with regards to issues such as domestic violence that are extremely normalised and underreported in Brazil.

6.4 Policy Implications and Final Conclusions

Nevertheless, this research still identifies important implications for policy. As this dissertation argues that violence in Brazil is characterised by inequality, it calls for the use of redistributive economic policy as a long-term preventative measure of violence. It highlights the need for an integrated approach to policy, whereby social issues are intertwined with economic issues. Most importantly, it reveals that the mere existence of political and economic structures of inequality promotes social suffering. The need to challenge this is not just a matter of justice and fairness, it is vital in the pursuit of a better society in which socially erosive forces are limited, and the well-being of all individuals is heightened.
Bibliography


Kasia Lipka

The trade-off between justice and economic efficiency is a key debate in welfare state economics. This study addresses whether a trade-off is inevitable and whether the financial crisis of 2008 has influenced this. An index measure of efficiency and justice is produced in order to conduct both cross-country and cross-time comparisons of performance across three European welfare states: Sweden, Germany and the United Kingdom, before and after the financial crisis. Detailed case studies are conducted to in order to explain the results in policy terms and to determine the conditions under which a trade-off may exist. The following findings are presented; 1) A trade-off between justice and efficiency is not inevitable, but it occurs in certain circumstances, 2) The existence of a trade-off is largely determined by the extensiveness of the tax system and the role of core societal values, 3) The financial crisis of 2008 has had a mixed impact on the existence of a trade-off in European welfare states.
1. Introduction

The welfare state describes a ‘variety of political practices and processes related to the arrangement of a social and economic order within a [state]’ (Schulz-Forberg, 2012, p. 1) and is charged with protecting the well-being of its citizens (Oxford English Dictionary, 2015). One of the main concerns in welfare state economics is whether the state can simultaneously pursue justice and economic efficiency, or whether it is subject to a trade-off between the two objectives (Barr, 1993). As the role of the welfare state has been increasingly scrutinised since the financial crisis in 2008, this recent period offers an interesting point in time to conduct such a study.

As will be discussed in Section 2, the literature does not specifically address the concept of justice in the context of a trade-off between social outcomes and efficiency. This gap in the literature provides the motivation for this study and shapes its purpose: to determine whether a trade-off between justice and efficiency is inevitable, and whether the financial crisis has had an impact on this. Section 3 provides the theoretical perspectives of the trade-off and defines the concepts of efficiency and justice as used in this study.

The study adopts an index measure approach modelled on the Human Development Index, which enables both cross-country and cross-time comparisons. Cross-country comparisons are arranged using Gosta Esping-Andersen’s typology of welfare states, which will be discussed further in Section 3, and cross-time comparisons are arranged around the focal point of the financial crisis. In addition, the study adopts a case-study approach in order to explain the results in policy terms. Section 4 describes and justifies this methodology.

The results in Section 5 present conflicting evidence that a trade-off between efficiency and justice is inevitable. In both the pre-crisis and post-crisis periods, as well as in the transition between the two periods, there were two comparisons indicating that a trade-off exists, and one comparison falsifying this. In order to address the conflicting evidence, Section 6 provides a detailed discussion of the results within each country in policy terms.

The paper will conclude that welfare states are not inevitably subject to a trade-off between justice and efficiency, but there are circumstances in which a trade-off will emerge. It will highlight the importance of two features in determining whether a trade-off exists. Firstly, an extensive and progressive tax system creates positive gains for both justice and efficiency,
reducing the trade-off between the two objectives (Busch, 2010). This is demonstrated by Sweden’s success in harmonising justice and efficiency in the pre-crisis period through an extensive tax system, negating the trade-off between the two objectives. Secondly, positive core societal values can encourage the extensive role of the state and improve its ability to avoid a trade-off between the two objectives. The widespread acceptance of taxation in Sweden and of the Kurzarbeit work reforms in Germany, in contrast to the hostile attitudes towards the welfare state in the UK, illustrates the influence of such core societal values.

Added to this, the paper describes the specific impacts of the financial crisis on the countries presented in the study. Sweden has become subject to a trade-off since the crisis due to a fall in the extensiveness of its tax system and a fall in the generosity of unemployment benefits. Germany has become less subject to a trade-off since the crisis due to an increased extensiveness in the tax system and the positive role of core societal values. The UK has continued to be subject to a trade-off since the crisis due to the limited role of its tax system and the hostility towards the role of the welfare state instilled in core societal values.

2. Literature review

There is a key limitation to presenting a comprehensive literature review for this topic. Namely, that the concept of justice has not been specifically addressed in the context of a trade-off between efficiency and social outcomes. Thus, the purpose of this section is to present the literature addressing the various aspects of the topic, to highlight the gaps and to explain how this paper will contribute to the topic.

Following the onset of the crisis European welfare states helped to cushion the blow from the initial economic shocks. This occurred through the functioning of automatic stabilisers, which simultaneously reduced the impact on income inequalities and offered a stabilisation effect to aggregate demand (Andersen, 2012; Basso et al., 2012). However, the unsustainability of public finances has since required the implementation of austerity measures in European states. The literature discusses the likely impact of such measures on both efficiency and social outcomes.

According to the social investment approach, the fall in social spending will cause a fall in labour productivity, due to the notion that social spending is a productive factor (Hemerijck and Vandenbroucke, 2012). Hughes and Saleen (2012) provide evidence to support this, finding that persistently low labour productivity has been a feature common across many European states.
following the financial crisis, with the UK’s performance standing out as particularly poor. Oulton and Sebastia-Barriel (2013) attribute the UK’s poor performance to the negative impact of the financial crisis on the productive capacity of the UK economy, which has restricted its ability to replicate the strong productivity growth in the pre-crisis years. This verdict suggests that the financial crisis could not only have a short-run impact, but also a long-run effect on labour productivity growth. Despite this, Hemerijck and Vandenbroucke (2012) propose that Sweden’s “productivist” social policy has encouraged the economy’s competitive strengths, enabling Sweden to maintain productivity levels after the crisis. The maintenance of these strengths will be dependent on the state’s ability to enhance the productive capacity of its economy in the context of austerity packages (Outlon and Sebastia-Barriel, 2013).

As well as the negative impact on labour productivity growth, Oulton and Sebastia-Barriel (2013) predict that the financial crisis will have a long-run effect on raising unemployment, which is a trend they observed after most financial crises. Guichard and Rusticelli (2010) support this prediction and propose that the hysteresis effects associated with a rise in long-term unemployment could cause a persistent rise in structural unemployment. Despite this, Guichard and Rusticelli (2010) argue that differences in institutional arrangements and policies may result in different labour market outcomes. For example, labour market reforms in Germany have facilitated labour market adjustment since the crash, helping to reduce unemployment rather than allowing it to increase (Contessi and Li, 2013). Contrastingly in Sweden, which is characterised by a high level of unemployment benefit generosity (Stovicek and Turrini, 2012), the financial crisis is likely to result in a rise in unemployment due to the weak incentive structures in the labour market and the ‘safety net’ function of the welfare state (Oulton and Sebastia-Barriel, 2013; Lister, 2009).

Added to the efficiency losses associated with the financial crisis and the subsequent fiscal consolidation in Europe, Andersen (2012) predicts that the austerity measures will also result in considerable social costs. A fall in social protection through austerity packages is particularly expected to increase the burden of poverty, social instability and economic inequalities (Hemerijck and Vandenbroucke, 2012). Social costs are expected to be greater in the UK and Germany, due to the implementation of relatively large austerity packages, than in Sweden, where fewer cutbacks are expected and the welfare state provides an extensive ‘safety net’ function (Vis et al., 2011; Lister, 2009).
Thus, the literature predicts that the financial crisis and subsequent fiscal tightening will have negative implications on both efficiency and social outcomes. However, the implications are likely to vary across countries, due to the differences in institutional arrangements (Guichard and Rusticelli, 2010). In the UK, the vulnerability of public finances has resulted in the implementation of extensive austerity packages. Diamond and Lodge (2013) predict that the reduction in social investment could reach an undesirably low equilibrium, restricting both social outcomes and efficiency. In Germany, austerity packages are expected to result in increased social costs. However, labour market reforms have thus far improved employment efficiencies, reducing the negative impact of the financial crisis on the level of unemployment. In Sweden, the generosity of unemployment benefits is expected to cause an increase in unemployment. However, Hemerijck and Vandenbroucke, (2012) highlight the gains in Sweden from capacitating welfare provision, which they argue has the potential to promote gains in both efficiency and social outcomes.

This section has presented the relevant literature and has highlighted the limitations of the literature, which provide a motivation and shape the purpose of this paper; to determine the impact of the financial crisis on the trade-off between efficiency and justice.

3. Theoretical Perspective

The purpose of this section is to provide the theoretical perspectives on the key issues examined in this paper. It will define the concepts of efficiency and justice and explore the trade-off between them, which is conceptualised as a key issue in welfare state economics (Barr, 1993). It will also consider the limited theory on the impact of financial crises on this trade-off.

The concept of justice is widely debated and has been defined in a variety of different ways. Arguably the most comprehensive available theory of justice is that conceptualised by John Rawls (Buchanan, 1982). In his ‘Theory of Justice’ (1971) he proposes that social and economic inequalities should be regulated by two principles. The first principle concerns basic liberties and human rights, which have priority over economic and social advantages. In Tungodden’s (1996) interpretation of Rawls’ theory it is proposed that poverty, which reflects severe injustice, is defined within the context of the first principle and thus poverty prevention is addressed within this study.

The second principle is a combination of fair equality of opportunity and the Difference Principle. Fair equality of opportunity requires that there exists the opportunity for all to acquire the skills
necessary to reach all positions in society, which can be measured by the degree of equitable education (Rawls, 1971). The Difference Principle marks the departure of Rawls’ theory from the traditional focus on equality. Rawls argues that, if the absolute priority in society is equality, the Pareto principle is violated (ibid.). He therefore proposes that inequalities can be permitted so long as they are arranged in such a way that they benefit the worst off in society (Tungodden, 1996). In modern capitalist societies, the system of direct taxation on personal incomes can facilitate this arrangement. The more extensive and progressive this system, the higher will be the satisfaction of the Difference Principle. This study will measure this through the level of average and top tax income tax rates, as well as the level of tax revenue as a percentage of GDP.

The concept of economic efficiency can be broken down into dynamic efficiency and static Pareto efficiency. The growth in labour productivity provides a useful measure of dynamic economic efficiency, incorporating several economic indicators and offering a dynamic measure of economic growth and competitiveness (Freeman, 2008). The measurement of Pareto efficiency is particularly relevant to this study due to its prominence in Rawls’ work and it can be measured by the level of unemployment, which represents one of the largest inefficiencies in capitalist societies. It is also important to consider the ‘natural rate’ of unemployment, as conceptualised by Friedman and Phelps as an equilibrium measure, as this gives an indication of long-run efficiency prospects (Blanchard and Katz, 1996). Weidner and Williams (2011) state that, although the ‘natural rate’ cannot be measured, its trend can be inferred from other information. They observe factors that could drive the ‘natural rate’ upwards; namely, an increase in the generosity of unemployment benefits and an increase in the rate of long-term unemployment. A rise in the natural rate represents a fall in Pareto efficiency, as would an increase in the mean rate of unemployment.

The theoretical notion behind the trade-off between justice and efficiency is that the generosity of the welfare state and the extensiveness of its taxation system, which promote redistribution and the satisfaction of the Difference Principle, hinder the productive capacity and efficient functioning of the economy (Taylor-Gooby et al., 2004).

This is illustrated clearly in the labour market. Generous unemployment benefits, which promote both principles of justice, can reduce labour market incentive structures by creating a ‘safety net’, whereby income levels and living standards are protected in the event of unemployment (Stovicek and Turrini, 2012). This can have adverse effects on both labour
productivity and the unemployment level. In terms of Pareto efficiency, a considerable amount of empirical evidence confirms the positive correlation between unemployment benefits and the level of unemployment (Messacar, 2014). In terms of dynamic efficiency, the reduction in employment incentives can result in inactivity traps, benefit dependence and falling employability, which negatively impacts growth in labour productivity (Stovicek and Turrini, 2012).

The conflicting impact of generous unemployment benefits on justice and efficiency thus illustrates a trade-off between the two objectives. However, the extent of this trade-off can vary. In his ‘Three Worlds of Welfare Capitalism’ (1990), Esping-Andersen forms a typology whereby welfare states are grouped into regime types, determined by the relative priorities attached to efficiency and justice, which in turn determine welfare state outcomes. The three regime-types conceptualised are Liberal, Corporatist and Social Democratic (Esping-Andersen, 1990). The priorities associated with these states can be used to predict the performance of welfare states in justice and efficiency measure; Liberal states are expected to rank the highest in the economic efficiency dimension and lowest in the justice dimension, Social Democratic states are expected to rank highest in the justice dimension and lowest in the efficiency dimension, and Corporatist states could be expected to rank in the middle in both dimensions. However, Esping-Andersen submits that most states are a hybrid of these typologies, and therefore the results are not as simple as described above. Table 1 summarises the priorities associated with each regime type and the examples of each regime to be used in this paper.

Table 1: Welfare State regimes

<table>
<thead>
<tr>
<th>Welfare state</th>
<th>Country</th>
<th>Priorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>UK</td>
<td>Promote economic performance and efficiencies</td>
</tr>
<tr>
<td>Corporatist</td>
<td>Germany</td>
<td>Promote social stability</td>
</tr>
<tr>
<td>Social Democratic</td>
<td>Sweden</td>
<td>Minimise poverty, promote income redistribution</td>
</tr>
</tbody>
</table>

Source(s): Adapted from Headey et al. (2000).

Many academics have confirmed Esping-Andersen’s typology, yet others have challenged and extended it (Ferragina and Seeleib-Kaiser, 2011). The strongest criticism is that typologies in general fail to ‘capture the complexity of different social arrangements’ (ibid., p. 598). Despite
this, Ferragina and Selleib-Kaiser’s literature review supports the typology proposed by Esping-
Andersen, arguing that it ‘provides an excellent starting point and heuristic device for empirical 
research’ (ibid., p. 597). It therefore provides a sound framework for this paper, as will be 
discussed in Section 4.

Financial crises are theorised as factors that instigate institutional and policy reform (Vis et al., 
2011). Applying Esping-Andersen’s typology, changes to economic fundamentals and 
institutions are expected to alter outcomes in terms of justice and efficiency, yet still in the 
context of a trade-off between the two objectives. However, two contesting views challenge 
this hypothesis. Financial and economic crises can intensify the demands on the welfare state 
and therefore may cause a simultaneous fall in performance in both objectives, falsifying the 
existence of a trade-off (Diamond and Lodge, 2013). Contrastingly, such crises may provide an 
opportunity for corrective state intervention, via the tax system, to simultaneously promote 
gains in economic efficiency and social justice, again challenging the existence of a trade-off 
(Busch, 2010).

The theoretical perspectives outlined in this section will be considered throughout the paper 
and will particularly contribute to the discussion of the results in policy terms in Section 6.

4. Methodology

The purpose of this section is to explain and justify the methodology used in this study.

The study computes an index in order to facilitate cross-country comparisons of performance in 
efficiency and justice. An index is a ‘composite measure that summarises and rank-orders 
several specific observations and represents some more-general dimension’ (Babbie, 2012, p. 
159). The index used in this paper is modelled on the Human Development Index (HDI), which 
has been a key element of the United Nations Human Development Reports since 1990 (Sagar 
and Najam, 1997). Although subject to criticism, the HDI is one of the most widely used 
measures of development (Wolff et al., 2011), and it thus provides a sound measure on which 
to model the index in this study.

Figure 1 provides a graphical illustration of the index structure. Firstly, relevant indicators are 
used to produce a score for the two components of efficiency (Appendix 1), which are then 
combined to produce an overall efficiency score. This process is repeated for justice; with
indicators compiled to produce scores for the two principles of justice before being combined to produce an overall score. These scores can be compared against each other as separate indices and then be combined to produce a composite index for efficiency and justice. It is important to have separate indices before the composite one in order to enable the study to compare the two dimensions and assess whether or not a trade-off between the two exists.

Figure 1: Calculation of the Efficiency-Justice Indices - Graphical representation

There are two steps to creating the index.
**Step 1. Creating indicator indices**

Using time series data, minimum and maximum values for each indicator are determined (Appendix 2). These values act as ‘goalposts’ so that an index between 0 and 1 can be calculated for each indicator (Human Development Report, 2013). The values chosen are the minimum and maximum values observed in the data sets used (Appendix 2).

Once these values have been determined, the value for each indicator is calculated using equation 1, with the exception of the ‘Unemployment rate’ and ‘At risk poverty/social exclusion’ indicators which are calculated using equation 2.

*Equations used to calculate indicator values*

\[
\text{Indicator value} = \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}} \quad (1)
\]

\[
\text{Indicator value} = 1 - \left( \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}} \right) \quad (2)
\]

**Step 2. Creating the sub-dimension, dimension and combined indices**

The sub-dimension indices are calculated by taking the geometric mean of the components within that sub-dimension. For example, the sub-dimension index for Dynamic efficiency is the geometric mean of the ‘Growth in GDP/hours worked’ indicator and the ‘Growth in labour utilisation’ indicator. This approach is then used to calculate the dimension indices. The geometric mean of the sub-dimension indices is calculated in order to generate the dimension indices. For example, the dimension index for efficiency is the geometric mean of the ‘Dynamic efficiency’ and ‘Pareto efficiency’ indices. Finally, the combined index is the geometric mean of the Justice and Efficiency dimension indices.

As with the HDI, equal weighting is applied across the sub-dimension indices, dimension indices and the combined index in this study. For example, firstly, the ‘Growth in GDP/hours worked’ and ‘Growth in labour utilisation’ indicators are given equal weighing in the ‘Dynamic efficiency’ index. Secondly, the ‘Dynamic efficiency’ index is given equal weighting with the ‘Pareto efficiency’ index within the ‘Efficiency’ index. Finally, the ‘Efficiency’ index is weighted equally with the ‘Justice’ index in the combined index. This does not mean that weighting is not applied, but rather indicates that weights are equal (Giovannini *et al.*, 2005).
The main reason for this approach lies in the purpose of the study, which is to determine whether or not a trade-off between efficiency and justice is inevitable. If the purpose were instead to determine how different preferences between justice and efficiency influence the index value, then experimenting with weightings would be necessary. In this study, however, the preferences are determined by the priorities of each welfare state and are therefore within the performance in terms of justice and efficiency. Thus, applying uneven weightings would distort, rather than enhance, results by favouring welfare states with certain preferences and would not contribute to the purpose of the study.

The results section of the paper conducts cross-country comparisons of the indices in order to determine whether or not a trade-off can be observed between efficiency and justice. To complement this, the paper also discusses the results in policy terms through a case study approach. This involves conducting an in-depth analysis of the topic through the study of a select few ‘cases’, whilst taking into consideration the relevant contextual factors (Mills et al., 2010). This approach enables the study to tackle the topic in an in-depth manner and to gain a real-world perspective on the relevant issues (Yin, 2009). One of the criticisms of case-based research is that it is characterised by a lack of selectivity, meaning that irrelevant details are included and can detract attention from the key issues (Siggelkow, 2007). This paper has addressed this through clearly defining the key concepts of the study in the theoretical perspective section, namely Rawls’ theory of justice and the concepts of dynamic and Pareto efficiency. The indicators used in the study have been selected to conform to these definitions to assure that irrelevant factors are not included.

Another criticism of the case-study approach is that a small number of cases may not fully capture all of the issues related to the topic or be a representative sample of the real world. This issue is tackled by carefully selecting the cases to be included in the study, in order that they are representative of welfare states in a wider context, rather than being confined to the specific cases in question. As discussed in section 3, Esping-Andersen’s typology of welfare states provides a useful framework for cross-country comparative studies. This study therefore uses three countries as cases, one from each typology, in order to identify whether a trade-off exists between justice and efficiency, and whether this varies between welfare state types.

The study covers two time periods, in order to determine whether there are changes in performance over time that indicate a trade-off between justice and efficiency. The two-time
periods chosen are separated by the financial crisis of 2008, which represents a significant challenge for welfare states and therefore offers an interesting point in time to base the comparisons on. The pre-crisis period covers the years 2001-2007 and the post crisis period covers the years 2009-2013.

5. Results

The results section is divided into three sub-sections. Section 5.1 presents the results in the pre-crisis period, 5.2 presents the results in the post-crisis period and 5.3 presents the changes in performance experienced by each country between the two periods. Each section considers whether the results provide evidence that a trade-off inevitably exists between efficiency and justice. The results are discussed in policy terms in Section 6.

5.1. Pre-crisis period

The summary results for the pre-crisis period are presented in Table 2 and Chart 1.

<table>
<thead>
<tr>
<th>Country</th>
<th>Efficiency Index</th>
<th>Justice Index</th>
<th>Combined Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>0.69</td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td>Germany</td>
<td>0.62</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>UK</td>
<td>0.71</td>
<td>0.54</td>
<td>0.62</td>
</tr>
</tbody>
</table>

In the pre-crisis period, Germany is plotted to the north-west of the UK (see Chart 1), scoring higher in the justice index but lower in the efficiency index, suggesting that a trade-off exists between justice and efficiency. Thus, a high efficiency score can only be achieved at the expense of a relatively low justice score, and vice versa.

However, Sweden’s performance challenges the conclusion of a trade-off. It scored similarly to the UK in the efficiency index yet achieved a much higher justice score, indicating that it did not need to sacrifice efficiency in order to achieve a higher justice score. Added to this, Sweden was plotted to the north-east of Germany, outperforming Germany in both dimensions, again indicating that it did not need to sacrifice performance in one dimension in order to have a higher score in the other. Both of the comparisons against Sweden’s performance therefore suggest that a trade-off is not inevitable, demonstrating that it is possible to improve justice without damaging efficiency, and vice versa.
5.2. Post-crisis period

The post-crisis summary results are presented in Table 3 and Chart 2.

Table 3: Post-crisis summary results

<table>
<thead>
<tr>
<th>Country</th>
<th>Efficiency Index</th>
<th>Justice Index</th>
<th>Combined Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>0.62</td>
<td>0.82</td>
<td>0.71</td>
</tr>
<tr>
<td>Germany</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>UK</td>
<td>0.63</td>
<td>0.57</td>
<td>0.60</td>
</tr>
</tbody>
</table>
The results in the post-crisis period lead to the following conclusion: Sweden performed to the North-East of Germany, suggesting that a trade-off exists, i.e. it had to sacrifice some efficiency in order to achieve a higher score in the justice dimension. However, the UK’s performance challenges this conclusion. It scored almost the same as Sweden in efficiency but considerably lower in justice, reflecting Sweden’s ability to have a higher justice score without sacrificing too much efficiency. Added to this, Germany performed to the north-east of UK, outperforming it in both dimensions. Both of the comparisons with the UK’s performance falsify the existence of a trade-off between justice and efficiency, similar to Sweden in the pre-crisis period.

5.3 Pre-crisis to post-crisis transition
The transition in performance from the Pre-crisis to Post-crisis period is presented in Chart 3. The transition in performance from the pre-crisis to the post-crisis period leads to similarly mixed conclusions. The UK moved in a north-westwards direction, improving its justice score at the expense of a lower efficiency score, replicating a movement along a trade-off path. However, the notion of a trade-off is falsified by both Sweden and Germany’s performance. Sweden moved in a south-westwards direction, experiencing a fall in the score in both dimensions which suggests it may be subject to policy failures rather than a trade-off.
between the two dimensions. Germany’s performance moved in the opposite direction; north-eastwards, becoming both more efficient and more just, again disproving the existence of a trade-off.

Chart 3: Pre-crisis to post-crisis transition

6. Discussion

The preceding results section presented conflicting evidence of an inevitable trade-off between efficiency and justice. In both the pre-crisis and post-crisis periods there exists evidence in one cross-country comparison that a trade-off did exist, yet in the other two cross-country comparisons this was not the case. The transition from the pre-crisis to post-crisis period for each country also leads to the same results; one example indicated that a trade-off existed, yet the other two examples falsified this. The purpose of this section is therefore to discuss the performance of each country in policy terms, by presenting the three case studies, and to address the mixed evidence in order to determine whether a trade-off exists since the crisis.

6.1. Case Study: Sweden

Table 4 shows the results for Sweden, both pre- and post-crisis. It presents the values for each indicator, sub-dimension and dimension, as well as the combined index value.
Table 4: Sweden: Breakdown of results, pre- and post-crisis

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Pre-crisis score</th>
<th>Post-crisis score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic efficiency</td>
<td>0.60</td>
<td>0.55</td>
</tr>
<tr>
<td>- Growth GDP/Hour worked</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td>- Growth Labour utilisation</td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td>Pareto efficiency</td>
<td>0.79</td>
<td>0.71</td>
</tr>
<tr>
<td>- Unemployment</td>
<td>0.79</td>
<td>0.71</td>
</tr>
<tr>
<td>Efficiency Index</td>
<td>0.69</td>
<td>0.62</td>
</tr>
<tr>
<td>First Principle Justice</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>- Poverty prevention</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>Second Principle Justice</td>
<td>0.74</td>
<td>0.71</td>
</tr>
<tr>
<td>- Equitable education</td>
<td>0.82</td>
<td>1.00</td>
</tr>
<tr>
<td>- Tax as % GDP</td>
<td>0.90</td>
<td>0.83</td>
</tr>
<tr>
<td>- Top income tax rate</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>- Average income tax rate</td>
<td>0.66</td>
<td>0.50</td>
</tr>
<tr>
<td>Justice Index</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>Combined Index</td>
<td>0.77</td>
<td>0.71</td>
</tr>
</tbody>
</table>

The breakdown of the justice dimension in Table 4 highlights two key characteristics of the Swedish welfare state. Firstly, its achievement of the highest possible score (1.00) in the ‘First Principle of Justice’ is a reflection of its success in tackling poverty, which is one of its most notable achievements and defining characteristics (Kangas and Palme, 2005). Secondly, its high score (0.90) in the ‘Tax as a percentage of GDP’ indicator reflects the strength of its redistributive functions. Lister (2009) also proposes that the core values engrained in Swedish society, most notably the widespread acceptance of taxation and the support for the extensive role of the welfare state, contribute to the strong performance in both of these indicators and the justice dimension in its entirety.

There is an on-going debate concerning the impact of Sweden’s strong performance in the justice dimension on economic efficiency. According to Lister (2009, p.244) Sweden has a perceived success in ‘marrying economic competitiveness with social justice’. However, the conflicting view is that the inadequate incentive structure in the labour market, the result of the generosity of the unemployment benefit system and a lack of investment in human capital, has

Table 4 shows mixed evidence of the impact of Sweden’s justice score on efficiency. Despite a relatively low dynamic efficiency score in the pre-crisis period, which was arguably a result of the poor incentive structure and lack of investment in human capital (Kangas and Palme, 2005), Sweden achieved a high Pareto Efficiency score due to its relatively low unemployment rate. Barth et al. (2014) propose that a high level of complementarity between non-market institutions and capitalist dynamics enables Sweden to escape the unemployment trap, particularly through effective wage co-ordination.

In the pre-crisis period therefore, as demonstrated by the results in the previous section, Sweden’s performance was not subject to a trade-off between justice and efficiency. The complementarity between capitalist dynamics and non-market institutions enabled Sweden to achieve a high justice score, without sacrificing too much efficiency. However, the disparity in the scores between the two dimensions (0.17), as well as the poor incentive structure in the labour market and lack of investment in human capital, may have future implications on performance.

Despite the strong performance in the pre-crisis period, in the post-crisis period Sweden’s performance dropped in both dimensions, causing a fall in the efficiency-justice equilibrium. It is important to examine the reasons behind this in order to determine the impact it has on the interaction between efficiency and justice.

The most significant change to the welfare state in the post-crisis period was the reduction in the benefit replacement rates, resulting in a considerable reduction in the generosity of the Swedish unemployment benefit system (Stovicek and Turrini, 2012). Table 4 shows the negative impact of this change on the justice dimension, with considerable reductions in the values of three justice indicators; poverty prevention, tax as a percentage of GDP and the average tax rate, representing a fall in the state’s redistributive properties.

Added to this, the expected efficiency gains of the reduced benefit generosity, and consequential improvement in the labour market incentive structure, were not realised in the post-crisis period. The transition of Sweden’s performance between the two periods was
therefore not subject to a trade-off. Instead, it was subject to a policy time lag, whereby the expected gains were not realised in the same period, causing a simultaneous drop in performance in both dimensions. This is arguably due to the influence of core values in Swedish society, particularly the social expectation that the welfare state acts as a ‘safety net’ (Lister, 2009), which did not immediately adapt to the policy change.

In terms of future predictions, the sustainability of Sweden’s high efficiency-justice equilibrium is largely dependent on maintaining the extensive role of the tax system (Kangas and Palme, 2005). Thus, if the tax system becomes less prominent in Swedish society a trade-off between efficiency and justice will likely emerge. There are early signs of this in the post-crisis period, with the comparative performance of Sweden and Germany indicating a trade-off exists.

6.2. Case Study: Germany

Table 5 shows the results for Germany, both pre- and post-crisis. It presents the values for each indicator, sub-dimension and dimension, as well as the combined index value.

Table 5: Germany: Breakdown of results, pre- and post-crisis

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Pre-crisis score</th>
<th>Post-crisis score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamic efficiency</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Growth GDP/Hour worked</td>
<td>0.54</td>
<td>0.50</td>
</tr>
<tr>
<td>- Growth Labour utilisation</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Pareto efficiency</strong></td>
<td>0.67</td>
<td>0.79</td>
</tr>
<tr>
<td>- Unemployment</td>
<td>0.67</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Efficiency Index</strong></td>
<td><strong>0.62</strong></td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>First Principle Justice</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Poverty prevention</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Second Principle Justice</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Equitable education</td>
<td>0.48</td>
<td>0.59</td>
</tr>
<tr>
<td>- Tax as % GDP</td>
<td>0.61</td>
<td>0.65</td>
</tr>
<tr>
<td>- Top income tax rate</td>
<td>0.50</td>
<td>0.49</td>
</tr>
<tr>
<td>- Average income tax rate</td>
<td>0.59</td>
<td>0.53</td>
</tr>
<tr>
<td><strong>Justice Index</strong></td>
<td><strong>0.64</strong></td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>Combined Index</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.67</strong></td>
</tr>
</tbody>
</table>
Germany scored poorly in the efficiency dimension in the pre-crisis period, reflecting its reputation as the ‘sick man of Europe’ in the early 2000s, whereby it experienced deteriorating competitiveness (Hallerberg, 2013). Hassel (2010) attributes a considerable proportion of this poor economic performance to state intervention, which hindered economic performance particularly through tight labour market regulation and generous unemployment support. However, its similarly poor performance in the justice dimension challenges the existence of a trade-off between the two dimensions when compared to Sweden, which simultaneously outperformed it in both dimensions. Despite this, there is evidence that a trade-off exists when Germany’s performance is compared to the UK’s, with Germany performing better in the justice dimension and worse in the efficiency dimension compared to the UK.

Although the pre-crisis period presented mixed evidence of a trade-off, the transition period implied that Germany was not subject to a trade-off, as it was able to simultaneously improve its performance in both dimensions.

The strength of Germany’s economic performance during and since the financial crisis has belied its reputation as the ‘sick man of Europe’. As illustrated in Table 5, Germany experienced a large improvement in Pareto efficiency, due to considerable reductions in the unemployment rate, which fell each year between 2009 and 2013. A slight reduction in the dynamic efficiency score in the post-crisis period reflects the need for investment in human capital and education in order to maintain and boost productivity growth (Busch, 2010). However, the overall efficiency score increased from 0.62 to 0.67 in the post-crisis period, despite the economic challenges posed by the financial crisis.

Many academics attribute Germany’s recent economic success to two Government labour market incentives; the Hartz reforms and the Kurzarbeit work programme. The Hartz reforms, gradually implemented between 2003 and 2005, were a set of labour market reforms that facilitated the modification of labour market policies in a comprehensive strategy (Jacobi and Kluve, 2006). The reforms reduced the generosity of unemployment benefits, significantly improving labour market incentive structures and contributing to the constant reduction in the annual unemployment rate since 2005, with the exception of a marginal increase in 2008. Alongside the Hartz reforms, the long-standing Kurzarbeit work programme, which was extended during the crisis, is ‘credited with helping Germany’s labour market adjustment’ during the financial crisis (Contessi and Li, 2013, p.1). Under the programme, the Government
compensates a percentage of the reduction in employees’ net earnings if wages and working hours are reduced during the downturn, in order to prevent the inevitable increase in unemployment. The widespread acceptance of the programme by workers and firms, due to its mutual beneficial nature, along with the strong government support and funding for the programme contributed to its success in reducing unemployment following the crisis (Contessi and Li, 2013).

Adding to the reduction in average annual unemployment rates since the crisis, Germany has also experienced a reduction in long-term unemployment rates and the generosity of employment benefits. According to Weidner and Williams (2011), both of these patterns indicate that the ‘natural rate’ of unemployment is falling. A fall in the ‘natural rate’ suggests the labour market reforms will have a longstanding positive impact on the unemployment rate, resulting in the likely improvement of Germany’s efficiency score over time.

The simultaneous improvement in the justice and efficiency dimensions has two possible reasons. Firstly, it is possible that there exists a high degree of complementarity between the welfare state’s redistributive functions and capitalist dynamics in Germany. Initiatives such as the Kurzarbeit work programme offer examples of this, whereby the active role of the welfare state through the tax system simultaneously promotes justice through its redistributive functions, and efficiency, by helping to maintain employment and productivity levels.

Secondly, the starting point at a relatively balanced low efficiency-justice equilibrium, with a combined score of 0.63 and a difference of 0.02 between the two dimensions, provides a greater capacity for the scores in both dimensions to simultaneously increase than if the combined score had been higher or if the difference between the two had been greater. Thus, it appears that Germany was not subject to a trade-off in the transition between the two the periods.

Despite the inherent lack of a trade-off in the transition between the two periods, Germany’s post-crisis performance reflects the existence of a trade-off when compared to Sweden’s performance, replicating the trade-off it experienced when compared to the UK in the pre-crisis period. Although it did not face a trade-off in the transition period, the fact that it appears to be subject to a trade-off both pre-crisis and post-crisis when compared to different countries suggests that it is subject to a trade-off at different levels of efficiency-justice equilibrium. However, the balanced nature of Germany’s performance, whereby the scores in both
dimensions were similar, could have positive future implications, as demonstrated by the lack of a trade-off in the transition between the two periods. For example, Hassel (2010) argues that the introduction of more liberal labour market policies occurred within the established conservative framework of institutions, and therefore the changes will not necessarily produce similar outcomes as in more liberal countries, such as causing a fall in the level of justice, reflecting a trade-off.

6.3. Case Study: United Kingdom

Table 6 shows the results for the UK, both pre- and post-crisis. It presents the values for each indicator, sub-dimension and dimension, as well as the combined index value.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Pre-crisis score</th>
<th>Post-crisis score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic efficiency</td>
<td>0.60</td>
<td>0.54</td>
</tr>
<tr>
<td>- Growth GDP/ Hour worked</td>
<td>0.57</td>
<td>0.46</td>
</tr>
<tr>
<td>- Growth Labour utilisation</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>Pareto efficiency</td>
<td>0.84</td>
<td>0.73</td>
</tr>
<tr>
<td>- Unemployment</td>
<td>0.84</td>
<td>0.73</td>
</tr>
<tr>
<td>Efficiency Index</td>
<td>0.71</td>
<td>0.63</td>
</tr>
<tr>
<td>First Principle Justice</td>
<td>0.69</td>
<td>0.64</td>
</tr>
<tr>
<td>- Poverty prevention</td>
<td>0.69</td>
<td>0.64</td>
</tr>
<tr>
<td>Second Principle Justice</td>
<td>0.43</td>
<td>0.50</td>
</tr>
<tr>
<td>- Equitable education</td>
<td>0.30</td>
<td>0.51</td>
</tr>
<tr>
<td>- Tax as % GDP</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>- Top income tax rate</td>
<td>0.39</td>
<td>0.49</td>
</tr>
<tr>
<td>- Average income tax rate</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>Justice Index</td>
<td>0.54</td>
<td>0.57</td>
</tr>
<tr>
<td>Combined Index</td>
<td>0.62</td>
<td>0.60</td>
</tr>
</tbody>
</table>

As discussed in the results section, the UK’s performance in the pre-crisis period displays evidence of a trade-off when compared to Germany’s performance, due to its achievement of a high efficiency score at the expense of a lower justice score. The low generosity of the UK’s welfare state and the regressive nature of its tax and benefit system hindered the effectiveness of the state’s redistributive functions, resulting in the low justice score in pre-crisis period (Diamond and Lodge, 2013). However, the limited generosity of the state enabled capitalist
dynamics to dominate markets and promote economic efficiencies, resulting in the high efficiency score in the pre-crisis period. The priority attached to economic efficiency, as proposed by Esping-Andersen as a characteristic of a liberal state, in the context of a trade-off therefore explains the relative performance in each dimension.

The transition of the UK’s performance between the two periods supports the conclusion that it is subject to a trade-off between justice and efficiency. Despite the welfare state’s priority of economic efficiency, the UK experienced a considerable drop in performance in the efficiency dimension in the post-crisis period. The decline in Pareto efficiency is a reflection of an increase in the average rate of unemployment from 5% in the period 2001-2007 to 7.9% in the period 2009-2013. However, the declining level of benefit generosity over the period could be an indication that the natural rate may begin to fall, as suggested by Weidner and Williams (2011), and the increase in unemployment is instead an increase in cyclical employment due to the economic downturn.

More importantly, the UK experienced a decline in dynamic efficiency due to a fall in the annual growth in labour productivity, which was negative for three of the five years in the post-crisis period. Labour productivity declined with economic growth after the financial crisis, but failed to recover in line with the recovery in economic growth (Oulton and Sebastia-Barriel, 2013). This is arguably the result of the negative impact of the financial crisis on the productive capacity of the economy, which has restricted the potential growth in labour productivity (ibid., 2013). This reduces the likelihood of the UK returning to the previous productivity levels and is likely to have a negative long run impact on labour productivity growth and therefore on dynamic efficiency (Oulton and Sebastia-Barriel, 2013).

The UK’s performance in the justice dimension moved in the opposite direction between the periods. A reduction in both the proportion of early school leavers and in the impact of socioeconomic factors on educational performance resulted in a considerable improvement in the UK’s score in the equitable education indicator (Schraad-Tischler and Kroll, 2014). Added to this, the UK’s top income tax rate, applicable to income over £150,000, was increased from 40% to 50% in 2010, before falling to 45% in 2013 (HMRC, 2014), having a positive impact on the top tax rate indicator. Together, the improvement in the equitable education and top tax rate indicators promoted a marginal improvement in the Justice dimension score.
Despite the evidence that the UK faced a trade-off between justice and efficiency in the pre-crisis and transition periods, there is evidence to suggest that a trade-off does not exist in the post-crisis period. Germany outperformed the UK in both dimensions, suggesting that there is capacity for the UK to simultaneously improve its performance in both dimensions. As proposed by Busch (2010) this could occur through corrective state intervention via the tax system. In terms of efficiency, an increase in investment in human capital, funded through the tax system, could increase the productive capacity of the economy and encourage future productivity growth. In terms of justice, an increase in tax levels to support the investment in human capital would likely improve the redistributive functions of the tax system, increasing the score in the justice dimension. However, there are early signs that the UK is not set to utilise this spare capacity, particularly due to a fall in the average rate of tax and a fall in the generosity of unemployment benefits, which does not bode well for performance in the justice dimension.

6.4. Addressing the conflicting evidence

There was evidence in the pre-crisis and post-crisis periods, as well as in the transition between the two periods, to suggest that a trade-off does not exist. However, each period also provided evidence to contradict this. A trade-off is therefore not inevitable, but it can occur under some circumstances. This section will consider the points discussed in the case studies to determine the circumstances under which a trade-off is more likely to occur and to determine whether a trade-off exists since the crisis.

The case studies highlighted the importance of the extensiveness of the tax system and the role of core values in determining performance in terms of justice and efficiency and the trade-off between the two.

As proposed by Busch (2010), corrective state intervention through the tax system can promote gains in both dimensions. In terms of justice, the more extensive and progressive the tax system, the higher will be the satisfaction of the Difference Principle due to the effective structuring of inequalities. In terms of efficiency, public investment in human capital encourages labour productivity growth and higher rates of employment (Busch, 2010). Sweden's pre-crisis performance offers a perfect example of this. Its success in harmonising justice and economic efficiencies through an extensive tax system resulted in the negation of a trade-off between the two objectives. Since the crisis, a fall in the extensiveness of the tax system and a fall in the generosity of unemployment benefits have resulted in the emergence of a trade-off, which will
likely increase if the extensiveness of the tax system continues to decline (Kangas and Palme, 2005).

Core societal values had a role to play in the performances in each country. In Sweden, widespread acceptance of taxation and support for the extensive role of the welfare state contributed to strong performance in both dimensions in the pre-crisis period (Lister, 2009). Likewise, in Germany the widespread acceptance of the Kurzarbeit programme in the transition between the two periods contributed to its success in reducing unemployment following the crisis (Contessi and Li, 2013). In contrast, hostile public attitudes towards the welfare state in the UK hindered its ability to simultaneously improve both objectives in the transition between the two periods through an extensive tax system and productive public spending, resulting in a trade-off between the two objectives.

7. Conclusion

This paper has disproved the inexorableness of a trade-off between justice and efficiency, which is defined as a key issue in welfare state economics (Barr, 1993), and uncovered the impact of the financial crisis on the trade-off. It has done so by producing an index measure of efficiency and justice in order to conduct both cross-country and cross-time comparisons of performance across three European welfare states: the UK, Germany and Sweden before and after the financial crisis.

The pre-crisis, post-crisis and transition periods presented conflicting evidence with regards to the existence of a trade-off between efficiency and justice. In each case, there were two examples indicating a trade-off and one example challenging this. The discussion of the relevant policy variables and contextual factors in each country offered explanations for this conflicting evidence. It is evident that a trade-off is not inevitable and that both the extensiveness of the taxation system and core societal values play key roles in negating the trade-off.

In Sweden, the pre-crisis period demonstrated the importance of the extensive tax system and positive core values in harmonising the objectives of justice and efficiency and negating a trade-off (Lister, 2009). However, a trade-off has emerged since the crisis due to a decline in the extensiveness of the tax system and a fall in the generosity of unemployment benefits. In Germany, the negation of a trade-off in the transition between the periods is a reflection of positive core societal values and an increased extensiveness of the tax system. Despite evidence
that a trade-off exists in the post-crisis period when compared to Sweden’s performance, the simultaneous improvement of both objectives in the transition period and the lack of a trade-off when compared to the UK’s performance suggests that Germany has become less subject to a trade-off since the financial crisis. If the tax system is extended further it is likely that Germany will replicate Sweden’s success in the pre-crisis period in harmonising the two objectives. In the UK, comparison with both Germany and Sweden in the post-crisis period suggests that there is capacity for the UK to simultaneously improve its performance in both dimensions, as a trade-off does not appear to exist. However, the limited role of the tax system and welfare state in the UK, which is exacerbated by hostile public attitudes towards the state, hinders its ability to pursue both objectives simultaneously.

The following conclusions are therefore apparent: Sweden has become subject to a trade-off since the crisis due to a fall in the extensiveness of its tax system and a fall in the generosity of unemployment benefits. Germany has become less subject to a trade-off since the crisis due to an increased extensiveness in the tax system and the positive role of core societal values. The UK has continued to be subject to a trade-off since the crisis due to the limited role of its tax system and the hostility towards the role of the welfare state, instilled in core societal values.

This paper has therefore disproved the inevitably of a trade-off between efficiency and justice, described the circumstances in which a trade-off might occur, and presented the varied impact of the financial crisis on the trade-off in three European welfare states.
Bibliography


166


## Appendix 1 – Indicator details: Definitions, data sources and time periods

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Definition/Calculation</th>
<th>Data source</th>
<th>Pre-crisis date</th>
<th>Post-crisis date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth in GDP/Hour worked</td>
<td>Average annual growth rate in GDP/Hour worked</td>
<td>OECD</td>
<td>2001-2007</td>
<td>2009-2013</td>
</tr>
<tr>
<td>Growth in labour utilisation</td>
<td>Average annual growth rate in labour utilisation</td>
<td>OECD</td>
<td>2001-2007</td>
<td>2009-2013</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>The average annual unemployment rate of people between 15 and 64 years of age</td>
<td>OECD</td>
<td>2001-2007</td>
<td>2009-2013</td>
</tr>
<tr>
<td>Poverty prevention</td>
<td>Population experiencing severe material deprivation, income poverty or at risk of poverty/social exclusion</td>
<td>SIM</td>
<td>2008</td>
<td>2014</td>
</tr>
<tr>
<td>Equitable education</td>
<td>Includes education policies, education expenditure, early school leavers and impact of socioeconomic factors on educational performance</td>
<td>SIM</td>
<td>2008</td>
<td>2014</td>
</tr>
<tr>
<td>Tax as % GDP</td>
<td>Average annual rate of total tax revenue as % GDP</td>
<td>OECD</td>
<td>2001-2007</td>
<td>2009-2013</td>
</tr>
<tr>
<td>Top tax rate</td>
<td>Average annual top statutory income tax rate</td>
<td>OECD</td>
<td>2001-2007</td>
<td>2009-2013</td>
</tr>
<tr>
<td>Average tax rate</td>
<td>Average annual average income tax rate</td>
<td>OECD</td>
<td>2001-2007</td>
<td>2009-2013</td>
</tr>
</tbody>
</table>

SIM – Social Inclusion Monitor Index Report (Schraad-Tischler and Kroll, 2014)
OECD – Organisation for Economic Co-operation and Development
## Appendix 2 – Indicator calculations: goal-post calculations and details

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unit</th>
<th>Minimum</th>
<th>Details</th>
<th>Maximum</th>
<th>Details</th>
<th>Dates</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax as % GDP</td>
<td>AAR</td>
<td>10.7</td>
<td>Turkey: 1971</td>
<td>49.5</td>
<td>Denmark: 2005</td>
<td>1980-2013</td>
<td>OECD</td>
</tr>
<tr>
<td>Average tax rate</td>
<td>AAR</td>
<td>0</td>
<td>Chile: 2000-2013</td>
<td>36.15</td>
<td>Denmark: 2012</td>
<td>2000-2013</td>
<td>OECD</td>
</tr>
</tbody>
</table>

OECD – Organisation for Economic Co-operation and Development  
EU – European Union  
AAG – Average annual growth rate  
AAR – Average annual rate  
PP – Percentage of population  
CD – Combined dimension
This paper seeks to assess the impact of the growth of prison labour in the US on its three main stakeholders: the prisoners; the firms employing the prisoners; and society as a whole. It will then explore whether the UK should follow a similar expansion, using in-depth interviews with eight important figures both in- and outside the UK prison industry as well as existing research into current practice. No previous study has compared prison labour in the UK with that in the US. Following analysis of the literature on prison literature in the US and interviews with relevant UK individuals, the paper concludes that prison labour has the potential to have a positive net effect on society.
1. Introduction

The aim of this paper is to analyse the effects of the growth of prison labour in the US and to evaluate the advantages and disadvantages of the UK following a similar expansion. In 2012, the then Secretary of State for Justice, Kenneth Clarke, declared that he and the government were aiming to double the quantity of prison labour in the UK by 2021 (Scherbel-Ball, 2012). Clarke believed that this would increase the annual revenue generated to £130m, therefore reducing the escalating costs of prisons (Wright, 2012). Furthermore, in 2012, the Department of Justice rebranded the old Prison Industries Unit as One3One solutions. The focus of this enterprise was to enable more prisoners to be in realistic work environments within the prison walls, while also reducing costs for prisons and taxpayers as a result.

Despite Clarke leaving his post in the 2012 Cabinet reshuffle, the aim of expanding prison labour still appears to be prevalent in the government’s thoughts. Training and work experience while in prison are beneficial for offenders seeking jobs once they are released, and as the current Secretary of State for Justice, Chris Grayling said (quoted in Blue Sky, 2014): “Clearly what is crucial to stopping re-offending is getting people into work.” This dissertation will analyse the productive potential of this aim, using research from the US and interviews with people both inside and outside the UK prison industry.

This paper will explore the impact of the growth of prison labour on three stakeholders in the US: the prisoners; the firms employing the prisoners; and society as a whole. The effectiveness of prison labour will be examined in terms of increasing employability, enabling easier reintroduction to the workforce. Similarly, the benefits prison labour affords to participating firms will also be analysed, including overall cost reduction, resulting from lower wages, leading to increased efficiency. Moreover, the societal impact of prison labour will be evaluated, looking at positives such as the substantial amount from prisoners’ wages which pays for restitution as well as room and board, and the potential negatives, including increased local unemployment as a result of unfair competition from the prison industry.

The methodology evaluates the advantages of in-depth interviews as well their limitations, and discusses why comparison between the US and UK is appropriate. The interview section outlines the views of those inside and outside the prison industry in the UK, with overall feedback being positive in terms of the potential of prison labour. This dissertation is part of the early stage of research in the
UK; however, the information gained through the interviews in particular supports many of the views expressed in the literature review on prison labour in the US.

2. Background

A number of sentencing and policy changes occurred in the US penal system in the 1970s. Phelps (2011) believes that these changes were the main cause of the significant increase in the correctional population, shown below in Figure 2.1. In 2013, there were an estimated 1,516,879 sentenced persons in state and federal prisons compared to just 314,000 in 1979 (Carson, 2014; BJS, 1980).

Figure 1: Level of Incarceration in the U.S. from 1925-2013

Mandatory minimum sentencing was one of the new laws and Judge Paul G. Cassell (2007) has described how it led to unduly harsh sentences, which caused the US taxpayer to suffer due to the increasing rate of incarceration and its high cost.

The Justice Improvement Act was created in 1979 and holds significant importance for this dissertation as it included the introduction of the Prison Industry Enhancement Certification Program (PIECP). This enabled private corporations to gain access to the prison labour market.

There are now specified statutory requirements to be met for a state to become eligible. These were finalised in 1999 and included the important clause that prisoners must be paid a prevailing wage which is ‘not less than that paid for work of a similar nature in the locality in which the work was...
performed’ (BJA, 1999 p.17002). However, the local or state prison industry can take up to 80 percent in deductions from gross wages for specified purposes including taxes, room and board payments and victims’ compensation. The number of businesses involved with this programme has increased steadily over the last 35 years (Auerbach, 2010). In the third quarter of 2014 PIECP employed 4,821 inmates within 200 operations (NCIA, 2014a).

The main employer of prison labour in the US is the Federal Prison Industries, better known under its trade name UNICOR. UNICOR’s production can only be sold to state agencies in order to ensure that private businesses do not face unfair competition in the open market. UNICOR employed 12,468 inmates as of September 30th 2014, 7% of the eligible prison population (FPI 2014). PIECP’s 200 operations employ 0.31% of the total prison population. Although these figures are small, the benefits of prison labour can still be evaluated and used to compare with the potential for prison labour in the UK. Both UNICOR and PIECP participants look to put prisoners in realistic work environments, enabling them to increase their employability to reduce recidivism (BJA, 2005; UNICOR, 2009).

3. Literature Review

The literature available on prison labour is not extensive. However, over the last 20 years there has been an increasing level of debate on the topic. As prison labour is a highly political issue, it can be difficult to find unbiased views on the subject. Bair (2004) believes he is one of the few impartial researchers, while Thompson’s (2012) paper is an example of research taking a biased view. It appears to show only the negatives of prison labour and as a result, only sources which support the negative view are included. In contrast, Reynolds’ (1997) analysis of prison labour for the National Centre for Policy Analysis focuses on its positives, emphasising the benefits it brings to firms, society and prisoners themselves. While researching this topic, the Urban Institute, an organisation which looks to ‘conduct sophisticated research to understand and solve real-world challenges in a rapidly urbanizing environment’, provided this dissertation with a number of unbiased and informative papers, from authors such as Atkinson and Rstad, exploring both the positives and negatives of prison labour and potential improvements (Urban Institute, 2014). These potential improvements can be an important indicator as to how the UK should implement and improve its prison labour programmes.

3.1 Prisoners: Positives

There are numerous potential benefits to prisoners who take part in prison labour. Therefore demand for these roles is high, making this workforce an attractive proposition for private enterprises. (Schwartzapfel, 2009). Kling (2002) describes how prison labour can benefit inmates, as they can help
to support their families while incarcerated and gain skills which will make them more employable on release, thus reducing the rate of recidivism. The importance of prisoners being able to contribute financially to their families is also stated by Solomon et al. (2004), as they benefit both emotionally and psychologically.

Improving prisoners’ skillsets is the main focus of most prison work schemes and many economists believe that these are successful (Saylor and Gaes, 1992; Kling, 2002; Solomon et al., 2004). Even though Bair (2004) states that, in its current form, prison labour is a form of slavery, he still accepts that a number of roles within prison can enable the prisoners to gain valuable job skills. These roles include welding, furniture manufacturing and sewing. Atkinson (2002) discusses how prison labour is beneficial for prisoners, especially when they are employed by the private sector, as they gain transferable and marketable skills. These employment opportunities occur through PIECP and have the added benefit of prisoners receiving higher wages and often direct contact with businesses.

PIECP creates many benefits to society as will be mentioned later, but is also, as evidenced in Smith et al.’s (2006) report, extremely important in achieving successful re-entry upon release. This report described how those involved with PIECP gained transferable skills which increased their employability, while the taxpayer also benefited from deductions from prisoners’ wages. However, in their analysis they describe how there are factors that affect the success of the programme, which are not included, weakening the validity of the results, but these factors are not defined.

Smith et al. (2006) are not the only researchers to have found evidence of prison labour reducing recidivism through the acquisition of skills. While analysing UNICOR, Saylor and Gaes (1992) discovered a statistically significant positive relationship between prison labour and lower recidivism. Their results found that not only were the prisoners in their study group 24.4% more likely to find a job post-release than those in the comparison group, but also that 12 months after release, only 6.6% of study offenders had had their parole licences revoked, compared to 10.1% of comparison offenders.

One issue with this study is that it was completed after the prisoners had been released for just a year. The long term benefits of working in a prison are, therefore, hard to estimate using this data.

Furthermore, Bair (2004) describes how Saylor and Gaes’ (1992) definition of recidivism is flawed. It looks at survival time, i.e. the time it takes for a released prisoner to commit another crime, rather than at the actual reduction of recidivism. Thus, when the paper states that survival time has
increased, this simply means that the length of time from the date of release up to the moment that another crime is committed by the released inmate has increased, not necessarily that he won’t become a recidivist due to the rehabilitative nature of prison work.

In addition, the prisoners in the control and experimental groups were carefully selected. The inmates in the experimental group were not allowed to have any disciplinary infractions and this led to selection bias (Maguire et al., 1988). Piehl (2003) agrees, stating that these inmates would have been the most likely to attain a job on release anyway, and also the least likely to commit another crime, reducing the validity of such a study.

Despite the issue of selection bias, additional evidence exists, which shows the potential benefits of prison labour. PRIDE (2013) Enterprises, a nationally recognised inmate training programme, certificated under PIECP, stated that inmates, who were released in 2010 and had worked for them for over six months, recidivated at a rate of just 10.55% in the two years following their release. The average rate of recidivism after three years for prisoners released in 2005 across 30 states was, in contrast, found to be 67.8%. This shows the positive impact of PIECP (Durose et al., 2014). In addition, 63% of their released inmates were placed in full time jobs in 2013 (PRIDE, 2013).

3.2 Prisoners: Negatives

Thompson (2012) believes that inmates do not gain higher employment prospects as a result of prison work, and that they are used by a number of companies to avoid health and safety regulations, enabling the firms to lower their production costs. Bair (2004) agrees that prisoners are often exploited and believes that this may actually make them unfit for capitalist society once they are released. This implies that prison labour does not increase its skills-set, confidence and employability, which could lead to a potentially high likelihood of recidivism.

Additionally, Bair (2004) and Atkinson and Rostad (2003) both discuss how the jobs available in prison are largely in old economy sectors, such as clothing and textiles, which are expected to decline due to technological improvements. Therefore, even if the prisoners gain skills in these sectors, they will have limited employment prospects upon release.

A further issue is the fact that the prisoners receive such low wages. Those working for UNICOR earn a maximum of $1.15 per hour, while those working within PIECP can have up to 80 percent of their wages deducted for room and board and restitution (UNICOR 2015, BJA 1999). Therefore prisoners
can end up receiving such a minimal income that they cannot provide assistance to their families or save for the future. In addition to this, Sloan (2010) describes how the majority of jurisdictions operating under PIECP have reduced wages for prison workers to the minimum wage, which is usually below the value paid for work of a similar nature outside the prison. They do this by creating long training programmes where the minimum wage is paid. A prisoner must work through this training period before they can be paid the prevailing wage.

In 2010, Auerbach (2010) compiled a summary of findings on the PIECP compliance assessments. Out of the 188 operations in existence at the time of the report, only 38 were assessed. In the findings, five of the fifteen jurisdictions assessed had problems with low wage levels for prison workers. These were resolved; however, as only 15 out of the 30 certified and active jurisdictions were assessed, this means that numerous other jurisdictions could have been paying their prison workers below the required level. Additionally, eight of the states and counties assessed use a training wage that starts at the minimum wage and increases over time, as mentioned by Sloan (2010). The length of time varies from 60 days in Idaho to 390 days in Mississippi (Auerbach, 2010).

3.3 Firms

As discussed earlier, prisoners can gain substantial benefits from prison labour, whether financially or by improving their skillset. Economists such as Atkinson (2002) believe that prisoners working with private enterprises, within PIECP, benefit more than those working for UNICOR. Similarly, Smith et al.’s (2006) report evidences how prison labour working with private enterprises is extremely important in achieving successful re-entry upon release. This section will examine the positive and negative aspects of prison labour, in relation to the firms involved.

There are three models in which a firm can coordinate with a prison in production of a good or service. The first is the customer model, where the state prison system uses the labour to produce a commodity which is supplied to a private enterprise for a given price (Bair, 2004). Here the prisoners are not in direct contact with the firm and therefore will gain the same transferable skills as if they were working for UNICOR.

The second model is the manpower model. Here the private enterprise may work directly with the prisoners to produce a product or service, using the prison’s equipment (Bair, 2004). This can reduce the firm’s costs further as they pay a lower price (or nothing) for the facility and the capital involved.
However, Bair (2004) describes how private enterprises in most cases use their own equipment, the employer model, but benefit from the prison’s space, often rent-free, and from reduced health and safety restrictions.

When a firm is working with prison labour then this state must be certificated by PIECP. Therefore the private enterprise must pay at least the minimum wage. However, the state agency can take up to 80 percent for taxes, room and board, family support and victims’ compensation (BJA, 1999). PIECP is different to both the UK prison industry and UNICOR, where prisoners can be paid significantly below the minimum wage and the prisons can still deduct a substantial percentage of this for fees mentioned above.

Given this, US firms may not benefit through substantially lower costs. Furthermore, Reynolds (1996) describes how prisoners often have lower productivity and a reduced range of skills. Additionally, prisons have increased security costs and if the private enterprise has to cover some of these, combined with the high turnover of workers, it may be difficult to use prison labour profitably. Conversely, according to the Enterprise Prison Institute (EPI) (2002) field survey, employers rate prison labour as a quality workforce. One employer in the survey explained how:

“Inmates learn that the success of our company depends on the satisfaction of our customers with our product. Quality, service and price have to meet expectations. Our futures are intertwined. They are justly proud of what we have accomplished.” (EPI, 2002 cited in Atkinson and Rostad, 2003, p.8) Bair (2004) believes that the work the prisoners are doing must enable them to improve their skills and therefore increase their employability. They should not be forced to do whatever work the warden demands of them. Jobs with private enterprises always face increased demand from prisoners; therefore increasing the number of firms working with prison labour appears to be beneficial. This is supported by Reynolds (1997) and Atkinson and Rostad (2003) who state that the government should look to increase the involvement of private companies with prison labour. The EPI (2002) survey found that 92% of the employers would recommend using prison labour to business associates and these employers stated that the quality and productivity of the prison labour was of higher importance than the lower cost.

If the firms involved in PIECP are able to train prisoners and increase productivity, while enjoying lower production costs due to free manufacturing space within the prison walls, then these companies can benefit. Absenteeism is not an issue and the prisoners are often so grateful for the distraction from
prison life that, in contrast to Reynolds’ (1996) view, their productivity is higher than average, as exemplified by the EPI survey. Furthermore, although it is not mentioned in the literature on prison labour in the US, firms who use prison labour and appear to be helping with the rehabilitative process, may market this as exhibiting an increased level of corporate social responsibility. This could improve their brand image and potentially increase their sales and market share.

3.4 Society: Positives

Reynolds (1996), Kling (2002) and Bair (2004) all highlight the importance of inmates being able to pay restitution for their crimes and reduce the cost of their incarceration to the taxpayer. The average cost of incarcerating a prisoner in the US in 2010 was $31,286 (Henrichson and Delaney, 2012). Therefore, if the prisoners contribute to the cost of their incarceration, this can further reduce the strain on the taxpayer and possibly create a potential Pareto improvement.

Figure 3.4.1 illustrates the allocation of inmates’ gross pay within PIECP. This table shows that in an 11 year period, over $270 million has been taken from prisoners’ pay and used in ways, which could create an aggregate welfare increase.

**Figure 3.4.1: Breakdown of the Allocation of Inmates’ Gross Wages**

![Wage Deductions from PIECP 2003-2014](image)

Source(s): NCIA, 2014b; BJA, 2004

Scott and Derrick (2006) discuss how prison labour can also benefit firms in the surrounding areas. Federal prisons will require inputs which are likely to be demanded from local businesses. UNICOR’s (2012) data supports this, as over $549 million of raw materials, supplies, equipment and services was procured from small businesses in 2012. Moreover, UNICOR is self-sustaining and operates at no cost to the taxpayer.
3.5 Society: Negatives

Although UNICOR does purchase a substantial quantity of raw materials and equipment from small businesses, there has recently been increasing publicity about prisons competing with private firms through the acquisition of government contracts. UNICOR can only supply to state agencies, but this does not mean they should be guaranteed these contracts. American Apparel, who bid for some of the same government contracts, has had to lay off 150 of their workers due to what they believe to be unfair competition from UNICOR, who can pay workers as low as 23 cents per hour (Fox, 2012). This is an example of where prison labour has had a negative impact on society as law-abiding workers have lost their jobs as a consequence. However, Kling (2002) believes that the consumer benefits in the form of lower prices.

Thompson (2012) strongly opposes prison labour and describes how it has taken real jobs from people in manufacturing. Atkinson and Rostad (2003) support this view, stating that inmate labour can result in companies losing contracts and some workers losing jobs, especially when it is conducted by prison enterprises like UNICOR.

Furthermore, PIECP does not have support from all researchers. Even though it is meant to ensure that non-inmate workers will not be displaced, Sloan (2010) found examples of where this is violated. One was the partnership of Lockhart Technologies and Wackenhut Correction (now known as GEO Group) which resulted in Lockhart transferring their production to the prison. Consequently they closed their business operation nearby which resulted in the termination of 150 jobs.

4. Methodology

In this section the advantages of conducting in-depth interviews will be discussed first. This will be followed by reasons for the comparison of the US and UK prison labour industries. Throughout this Section the limitations of both the interviews and the country comparison will be examined.

4.1 Interviews

Due to the lack of studies on prison labour in the UK, this dissertation includes an interview section, discussing the positives and negatives of prison labour with important figures who have been involved in the prison industry, as well as some who currently have not been influenced by it.

Boyce and Neale (2006) discuss how in-depth interviews are useful when you want to explore people’s opinions on a new idea or issue. As there is little literature on prison labour in the UK, the idea could
be seen as relatively new; therefore interviews appear to be an effective way of ascertaining people’s perspectives on this topic.

4.1.1 Type of Research

Merriam (2014) emphasises the difference between the categories of ‘basic research’ and ‘applied research’. ‘Basic research’ focuses on expanding one’s understanding of a topic or issue, while ‘applied research’ looks at improving the quality of practice of a particular discipline. Thus, as the interviews conducted for the present study are aimed at gaining an understanding of the opinions of those involved, or those, who could be involved with prison labour, they can be ascribed to the category of ‘basic research’.

Using the research gained from programmes in the US, further studies could explore ways of improving prison labour programmes in the UK - this would be ‘applied research’. The Howard League of Penal Reform (2011), a UK charity, has produced studies such as ‘Business behind bars: making real work in prison work’, which could be described as ‘applied research’ as it attempts to influence policy makers on the subject of prison work. Further studies could look at transferring knowledge gained from successful US programmes as recommended by this paper.

4.1.2 Limitations

Beale and Neal (2006) describe how interviews can be prone to bias. This implies, when relating to this dissertation, that if the respondent wants an expansion of prison labour, they may exaggerate the benefits and not mention the negatives. However, as this dissertation is unlikely to have any influence on future government policies, this bias is less likely to be present. Furthermore, Silverman (2013) describes how the benefit of doing interviews depends on the research problem itself. As there is limited statistical data on prison labour in the UK and limited expansive studies have been implemented, qualitative research appeared to be the most effective form for this dissertation.

One important limitation of in-depth interviews is that they are not generalisable (Beale and Neale, 2006). As the respondents in this dissertation were not chosen through random sampling methods and because the sample is small, the results gained from these interviews cannot be generalised for the opinion of society or those involved with the prison industry. However, these interviews do provide valuable information, as part of the first stage of research on the topic. This research will need to be expanded in the future through expansive qualitative and quantitative studies.
4.2 Why Use the US for Comparison Purposes?

Jones and Newburn (2002) explain how there appear to be a number of examples of policy transfer between the US and the UK in terms of crime control over the last 30 years, such as zero-tolerance policing and private sector involvement in corrections. Accordingly, the UK may look to successful US prison labour programmes in order to maximise the potential of prison labour in the UK.

4.2.1 One3One and UNICOR

The US prison industry has expanded substantially since the 1970s and the Federal Prison Industries adopted the trade name UNICOR in 1977 (UNICOR, 2009). Since its creation UNICOR has expanded its production, increasing revenues from $161 million in 1983 to $609 million in 2013 (Pederson, 2000; FPI, 2013). In comparison, the UK Prison Industries unit was rebranded as One3One solutions in 2012, 35 years after UNICOR was created (One3One, 2014b). There is substantially more research on prison labour in the US, which can be used as a reference for the UK prison industry.

As of 2014, UNICOR employed 12,468 inmates. The total number of inmates in federal prisons was 209,293 as of March 19th 2015; therefore, UNICOR currently employs roughly 6% of all inmates in federal prisons (BOP 2015). Previously, this paper referred to UNICOR employing 7% of the eligible prison population, and that is because a proportion of inmates is either under heightened security restrictions, which prevents them from working, or are mentally or physically incapable of doing the work.

One3One solutions employed 9,700 prisoners in 2013 out of a total prison population of 85,382 (One3One, 2014a; MOJ, 2013b). This means that roughly 11% of inmates in the UK are placed in real work environments within the prison walls. The similarity in the percentages of prisoners working for UNICOR and One3One makes the comparison of countries relevant.

However, one significant difference is that the majority of UNICOR’s produce is demanded by state agencies, whereas in the UK, One3One produces products for over 150 companies. This is due to the different political structures and laws of the two countries. Nevertheless, in both examples, the prison enterprises still need to produce at a sufficiently high quality and sufficiently low price to maintain demand.

4.2.2 PIECP

Although One3One Solutions does create products for businesses in the UK, it also incentivises UK firms to move their production to prisons. Therefore, a company such as LMB Textiles, mentioned
below in the Interview Section, has been able to move its production to a prison facility, rather than transfer it abroad. This is beneficial for the UK economy, as well as the prisoners and the firm.

The fact that UK prisons look to work in partnership with private enterprises, as well as just producing for them, is why comparison with PIECP in the US is useful. As discussed in the literature review, some economists believe that prisoners involved with PIECP gain more transferable skills than those working for UNICOR (Atkinson, 2002). This may be an indicator that the UK should look to involve more private enterprises directly with prison labour, rather than just use prison labour as a form of outsourcing production.

A significant difference between PIECP and the partnership of prisons and private enterprises in the UK, is that those firms in the US working with prison labour have to pay them at least the minimum wage (BJA, 1999). This means that UK firms are likely to gain an increased competitive advantage. Prisoner wages in the UK are determined by Prison Governors. They averaged £2 per hour in 2012 which may incentivise more UK firms to work with prison labour as it can substantially reduce costs (Scherbel-Ball, 2012).

Conversely, one benefit for society of PIECP is that millions of dollars of inmates’ wages have been used to contribute towards room and board and restitution (NCIA 2014a). Only firms suffer as a result of having to pay higher wages. Consequently, if the UK government is able to ensure that the productivity of UK prisoners and the potential benefits from corporate social responsibility are still high enough to incentivise private firms to partner with prisoners, it may be beneficial to society to increase the minimum wage for prisoners in the UK.

However, as this is currently not the case, comparing the benefits gained by US firms working with prison labour to the potential for UK firms may actually understate the advantages. The number of firms in PIECP has expanded since its creation, even though firms cannot pay low wages (Auerbach, 2010). Therefore the combination of the lower capital costs due to using prison facilities, with the productivity of the prisoners, must be enough to make using prison labour more profitable to certain US firms than producing elsewhere. If prison labour is profitable to US firms, despite having to pay at least the minimum wage, then the potential benefits for UK firms could be even higher.
5. Findings and Analysis

5.1 Interview Background

The majority of these interviews were done by telephone in 20-30 minute conversations. The topic was discussed with each interviewee. Following this conversation they received the set of focused questions for this section. They responded in writing, enabling direct quotations in this Section.

There is very little literature on the impact of prison labour in the UK, therefore this dissertation looks to use the views of those inside and outside the industry to uncover the positive or negative potential for prison labour in the UK. The section will focus on three areas which were also the main focus of the literature review. These are: the impact of prison labour on prisoners; the potential costs and benefits to firms using prison labour; and the impact of prison labour on society.

Table 5.1.1 shows the diverse range of interviewees, and this variety of opinion should enable the results to have reduced bias.

<table>
<thead>
<tr>
<th>Interviewee 1</th>
<th>Head of Reducing Reoffending at a UK prison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewee 2</td>
<td>Consultant for a rehabilitative scheme</td>
</tr>
<tr>
<td>Interviewee 3</td>
<td>Previous employer of prison labour</td>
</tr>
<tr>
<td>Interviewee 4</td>
<td>Economics broadcaster (produced a programme on prison labour)</td>
</tr>
<tr>
<td>Interviewee 5</td>
<td>Founder and Chief Executive of Key4Life (rehabilitative programme for prisoners)</td>
</tr>
<tr>
<td>Interviewee 6</td>
<td>Previous head of rehabilitation programmes for the Prison Service</td>
</tr>
<tr>
<td>Interviewee 7</td>
<td>Previous education worker in a prison</td>
</tr>
<tr>
<td>Interviewee 8</td>
<td>Head of HR for a manufacturing firm not using prison labour</td>
</tr>
</tbody>
</table>

5.2 The Impact of Prison Labour on Prisoners

As discussed in the literature review, studies in the US have shown that prison labour can have a positive effect in reducing recidivism and increasing the chance of a prisoner gaining a job on release, due to the skills they have gained from working in prison. Using this information it would be expected
that prison labour would have the same potential in the UK and the majority of my interviewees agreed.

When asked if prison labour would be beneficial Interviewee 1 said:

“Absolutely YES! Paid work provides each individual with a sense of worth and a sense of pride. Many offenders who may be disengaged from the mainstream community are afforded a level of self-belief that they are able to support themselves and their family. They are able to gain nationally recognised qualifications in specialist sectors, as well as the ability to build a CV to take forward to release.”

This shows that it is not only the skills gained by the prisoners, which are important, but also their psychological placement into society. This view was supported by Interviewee 6 who stated that: “doing paid work gives prisoners a sense of self-respect and self-worth” and Interviewee 2 who discussed the importance of the prisoners “gaining self-esteem” from working.

However, if the prisoners are only gaining a sense of pride and self-worth, and not transferable skills, which increase their employability, then prison work may not be that beneficial. Interviewee 7 thinks that “the value and skills gained by most educational activities surpassed those gained at work.”

Furthermore, interviewee 8 was “not convinced” by the benefits of prison labour. They stated how “[i]t could be perceived as menial work paid below the minimum wage.” This is also the view of the Howard League for Penal Reform (2008), who added that prison labour in the UK currently focuses more on soft skills such as time management and teamwork rather than transferable skills, which would increase employment prospects.

As part of the punishment for committing a crime, prisoners do not have to be paid the minimum wage. In fact, prisoners can be paid as little as £4 per week (MOJ, 2000). There is therefore potential for UK firms and prisons to take advantage of this, using prisoners in low-skilled manual tasks, which, in the US, has been described as slave labour by Bair (2004).

Interviewees 4 and 5 are worried about prisoners being exploited if wages are kept very low. Interviewee 4 believes that prisoners “should be paid the minimum wage per hour, from which should be deducted the cost of their board and lodging in prison and a contribution to security costs.” As mentioned in the literature review, this is required by law in the US when a private enterprise is
employing prison labour. The UK government may, therefore, want to look at implementing this as it would reduce the costs to prisons and hence the taxpayer. It is also likely to mean that prisoners could keep some of that higher wage, enabling them to help their families and/or save for life post-prison.

Interviewee 7 thinks that “gaining a job, quickly, is one of the key indicators that reduces recidivism.” However, this same interviewee feels that ‘the setup at the moment does not really contribute to this in any meaningful way.’

In a sample of 1,080 prisoners taken by the Ministry of Justice (2014), just 17% of them stated that they had received help to look for a job, training or education to benefit them post-release. Of the 83% that did not receive help, 51% said they would have liked help. Therefore, in this sample of 1,080 prisoners, around 184 received help, whereas an additional 453 who wanted help did not receive it. As gaining a job on release is so important, according to both prisoners, as described by Hopkins (2012), and by Interviewee 7 (previously an education officer in a prison), having a job while in prison, and having an interview guaranteed when released, could be extremely helpful in reducing the recidivism rate.

If a prisoner is doing paid work while in prison, the hope would be that they would gain transferable skills to give them an increased probability of employment post-release. However, evidence given by Brunton-Smith and Hopkins (2014), suggests that the majority of prisoners who gain work once they leave prison were working in that same job before conviction.

Despite these results, Interviewee 5 highlights the significance of the relationship between the firms employing prison labour and the prisoners themselves:

“Prisoners benefit from doing paid work so long as there is training, references written for them and a link back to the company employing them. It is important that the prisoners are not just seen as a commodity and are instead seen as potential full time employees of the company in the future.”

All of the interviewees believed that prison labour would benefit prisoners, whether by the ability to earn money, increased self-worth from the responsibility and respect gained from having and keeping a job, or simply reducing idleness. However, there is potential for prisoners to be exploited, working in menial tasks and earning very little, and there was some doubt among the interviewees about the likelihood of the prisoners developing transferable skills.
For prison labour to be successful in the UK, it is essential that prisoners are trained and treated with respect. Their punishment should be combined with the opportunity to rehabilitate. Interviewee 1 describes how “we should create more chances for people who have been in jail to make a positive contribution to the workforce.” This is supported by the opinion of Interviewee 6 who states that:

“Prisoners need to see and believe that it is possible to successfully reduce their risk of re-offending and that when they have, society won’t continue to judge them. I have employed ex-prisoners myself and they are often the most committed and loyal employees you could wish for.”

5.3 Do Private Firms Benefit from Using Prison Labour?

One3One (2014b), the government enterprise set up to attract firms to work with prison labour, describes it as “a workforce of motivated prisoners who are looking to repay society and build outstanding business relationships with you.” Interviewee 4 is also optimistic about prison labour and believes it is extremely beneficial for firms:

“They will find a group of committed employees inside prison, selected with help from staff for aptitude, and eager to work to escape the extreme boredom of prison life. On release, such prisoners can make good continuing employees and prove exceptionally loyal to the company which gave them a break from their past lives.”

In addition, Interviewee 5 feels that the positives of prison labour outweigh its potential negatives:

“It is a largely untapped workforce and if the firm looks to increase the employability of the prisoners, it can create a long term partnership between the firm, prisoners and prison which is beneficial for all. Firms have a corporate social responsibility and employing prison labour is a fantastic opportunity for these firms to boost their brand by increasing the employability of prisoners, giving them a second chance.”

Corporate social responsibility is mentioned by seven out of the eight interviewees and therefore appears to be an important potential positive of prison labour. If a firm is able to benefit from lower costs due to lower wages, while also improving their marketable image due to the rehabilitative benefits of employing prisoners, then this is likely to create a net benefit for the firm. Interviewee 1 summarises this:
“Global businesses have a corporate social responsibility to influence, support and provide initiatives to encourage the rehabilitation of prisoners. Research has shown that most ex-offenders are more committed and willing to do more than just the job. They are grateful for the opportunity to do something worthwhile rather than falling back on their former ways and circumstances.”

However, there are a number of potential negative effects of using prison labour; including lower skills, lower productivity and the increased security risk within the prison. This leads to Interviewee 2 stating that they “genuinely think that it is a high risk to employ prisoners.” Moreover, as Interviewee 7 describes, the whole process of employing prison labour may increase costs “due to bureaucracy”. Furthermore, Interviewee 5 believes that:

“the way the firm positions the employment is extremely important as to whether it is beneficial to them, especially in the long run. If firms are seen to be taking advantage of the prisoners, this could lead to a dilution of their brand’s reputation in the long run.”

Although there is very little academic literature on prison labour in the UK, there are a number of articles on the negative aspect of this practice. Speedy Hire reduced its workforce by 800 in 2010 and has since increased the size of its prison contract (Wright, 2012). It has therefore been accused of replacing its law-abiding workers with prison labour. This resulted in their public image suffering and also increased local unemployment, which has a negative impact on society. Furthermore, Becoming Green, a firm in Wales, has been employing more and more prisoners on ‘work experience’ and has subsequently been firing a number of their law-abiding workers (Malik, 2012). Although this dissertation is focusing on those prisoners working within the prison walls, this example could not be ignored as it shows the potential negative effect of expanding prison labour.

The above examples describe how a firm taking advantage of cheap prison labour can receive negative publicity which could reduce its profitability in the long run. However, these two firms appear to weigh the positives (lower wages and a motivated workforce) as greater than the negatives (reduced brand image due to negative publicity). It must be ensured that this is not the case for the majority of firms using prison labour, otherwise there could be a significant negative impact on society.

Despite these examples and the potential negative aspects of employing prison labour, the majority of the interviewees concluded that it was beneficial for firms. The Howard League for Penal Reform
supports this view and believes that the government should try to incentivise more private companies to operate within prisons. They believe that:

"the ultimate goal is an atmosphere where private business demand for real work in prison is high and businesses are able to approach governors at the prisons themselves" (Howard League, 2011 p.5).

An example showing the potential success of using prison labour is Norpro, the engineering company which Interviewee 4 mentioned. They were able to move their production facilities from India to HMP Altcourse, a prison near Liverpool. David Norburn, Norpro’s Managing Director, described why prison labour was so attractive to the company:

“I am a businessman first and foremost, and for me it had to make commercial sense to have my factory inside the prison. I didn't have a factory and this would involve no huge capital investment from me, other than the fees I could pay per unit produced...The results of working at Altcourse have been excellent.” (Norburn quoted in G4S, 2014)

5.4 The Overall Impact of Prison Labour on Society

If prison labour is benefitting the prisoners, then this is likely to have multiplied effects on society, potentially resulting in a Pareto improvement. Reduced recidivism leads to a lower number of people in prison. This means there is less strain on the taxpayer and improved conditions for the prisoners. According to the Ministry of Justice (2013a) the average cost of holding one person in prison for a year is £34,766, therefore every released prisoner who avoids a return to prison saves the taxpayer a significant amount.

The way prison labour is implemented is an extremely important factor in determining its overall impact on society, as discussed by Interviewee 5:

“Prison labour has the potential to reduce reconviction rates if the revenue prisons receive from prisoners’ wages is re-invested into rehabilitation schemes. If prison labour is just a win for the firm employing the prisoners, then this is not beneficial for society. If, however, prisoners are trained for when they leave prison and they receive an interview guarantee, then this can be very beneficial for society. Firms can reduce their costs, prisons can cover some of the costs of the prisoners while also helping them get rehabilitated, possibly stopping reoffending, and importantly the prisoners gain the skills which increase their employability, while they also have their deeper, more emotional problems dealt with.”
In summary, prison labour can: reduce the cost of prisons to the state, enabling the government to invest in other areas; increase the productivity of UK firms, allowing them to produce goods for lower prices, bringing benefits to the consumer and enabling the firm to compete more effectively abroad; and can facilitate the firm’s expansion, due to increased productivity, which in turn would result in increased employment. These are all reasons why prison labour could result in an aggregate welfare increase.

The issue of whether firms employing prison labour reduces employment of law-abiding citizens is an important one. As mentioned previously, companies such as Speedy Hire and Becoming Green appear to have replaced their law-abiding workers with prisoners. This has a negative impact on society as it increases unemployment, which not only reduces the tax revenue for the government, but also means that expenditure on benefits increases. However, Interviewee 7 believes that the “numbers involved with prison labour would be so low that they would not have a significant impact.”

Alternatively, prison labour can actually create numerous benefits for society and the economy when it enables firms to produce in the UK rather than abroad. Norpro’s experience showed how the reduced costs of producing in a prison enabled the firm to bring production back to the UK.

LMB Textiles is another firm which has benefited from the lower costs of prison labour enabling them to keep production in the UK, while their competitors moved their manufacturing abroad. Ross Barry, Business Development Manager, describes the benefits his firm has gained from using prison labour:

“Using prison workshops has helped LMB to expand and develop our business. It has allowed us to continue to work in the UK and given us extra capacity and the ability to expand quite quickly and easily should we need it.” (Barry quoted in One3One, 2014c)

In addition to this, Interviewee 3 had a positive outlook on the response of the population to prisoners working.

“It could create real benefits to the prisoners, the prison services and I am sure the general public would applaud that prisoners are contributing. I see work in prisons as an important tool in helping prisoners rehabilitate, improving their confidence and self-worth. We want people not to reoffend, and working in prison can only contribute to that goal.”
All eight of the interviewees believed that prison labour could benefit society overall, through increasing the employability of prisoners, reducing the costs of participating firms, and decreasing the strain on the taxpayer.

6. Limitations and Further Research

The main limitation of this study is the impossibility to generalise the results gained from the interview section due to the small sample size (Boyce and Neal, 2006). Furthermore, the paucity of literature on prison labour in both the US and the UK may limit the validity of the conclusions made.

This dissertation is part of the first stage of research on prison labour in the UK. No previous study has compared the US prison industry as a whole to that of the UK. However, despite the stated limitations, the results of this early research indicate that its conclusions could lead to changes in UK policy on prison labour. Further studies, both quantitative and qualitative, should focus on the impact of prison labour in the UK on recidivism, as well as explore the models as well as the benefits of the most productive prison labour programmes. The success of One3One Solutions must be evaluated and potential improvements could be gained from further research in the US.

7. Conclusion

The review of the literature on prison labour in the US gave the impression that an expansion of prison labour could lead to an aggregate welfare increase. In general, prisoners who join prison labour schemes gain transferable skills which make them more employable. This reduces recidivism as shown in the studies by Saylor and Gaes (1992), focusing on UNICOR, and Smith et al. (2006), analysing PIECP. Furthermore, since its introduction in 1979, PIECP has deducted over $415 million from prisoners’ wages for payments towards room and board, restitution and taxes (NCIA, 2014b). This has lowered the impact of incarceration on the taxpayer. In addition, the EPI (2012) survey has shown how the quality of work by prisoners has improved over recent years, resulting in firms benefiting not only from the lower cost of prison labour but also from high production quality. There were, however, a number of negatives outlined in this paper such as the occasional exploitation of prisoners by firms and prisons, as well as prisoners being trained in old industries, hence reducing their employment prospects upon release. In conclusion, both my research and the relevant studies show that prison labour appears to have had an overall positive impact in the US.

The UK interviews support the opinions held in the literature review on the US. All eight interviewees in this dissertation believed that prison labour could benefit society overall through increasing the
employability of prisoners, reducing the costs of participating firms, and decreasing the strain on the taxpayer. The interviews do, however, also identify the potential for prison labour to create negative externalities if firms are able to replace law-abiding workers with cheaper prison labour, or use prisoners in menial tasks which do not improve their skillset.

Therefore, the type of employment within the prison is critically important as to whether it is beneficial to society overall. If it involves training and improving the prisoners’ skillsets, this can have several positive effects for society, as it can enable the prisoner to gain a ‘good job’ which, according to both Interviewee 7 and Piehl (2003, p.6), is integral to reducing recidivism. The UK government must therefore ensure that One3One Solutions employs prisoners in roles that will help them gain employment on release. Moreover, increased partnerships between private enterprises and prisons should include the possibility of prisoners gaining employment with these firms on release.
Bibliography


How are rating agencies perceived following the US financial crisis? What were their main shortfalls? Insights from academics and professionals

Timothy Sapsford

This research project looks at how academics and professionals with economics or finance backgrounds perceive rating agencies and their shortfalls following the US financial crisis. The paper attempts to fill a gap in the present literature, namely, the lack of qualitative research depicting people’s perceptions on the subject matter. The researcher engaged in in-depth interviews to retrieve this data, where four central themes were identified: (i) loss of reputation, (ii) conflict of interest, (iii) complexity and lack of judgement and (iv) lack of competition in the industry. The sample believed that rating agencies’ reputation has been negatively affected. Nevertheless, at least some of the interviewees continue to see their overall function in financial markets as positive. The data was inconclusive in regard to whether a change in the business model affected the way rating agencies operate. It showed, however, that the sample believed that conflicting interests were inherent in the current system. The participants also showed their concerns about the complexity of the structured debt and the inability of raters to decipher the risk associated with them. The view that raters lost a grip on reality and lost their sense of judgement was also prevalent. Whilst the majority of participants believed the lack of competition in the industry affected rating agencies’ business operations, a general sense for the need for increased competition came about.
1. Introduction

1.1 Background

1.1.1 The US financial crisis and mortgage-backed securities

The financial crisis, which disrupted markets around the world between 2007 and 2009, saw the most severe depression since the 1930s, costing a total of $120,000 per US household (Atkinson et al., 2013). It is widely accepted that a bubble in real estate prices initiated the US financial crisis. This has been found to stem from the expansionary monetary policies put in place after the dot com bust (Allen and Carletti, 2010; Ackermann, 2008; Brunnermeier, 2009). In addition, other competing factors led many world economies to falter, (with notable exceptions; e.g. Australia) namely global imbalances, market deregulation and excessive leverage (Jickling, 2009).

Another reason was a shift in the banking model from an ‘originate and hold’ model to an ‘originate and distribute’ model. Banks would originally allocate loans to borrowers after the latter had met certain criteria of a rigorous screening process, and would subsequently bear the risks of any eventual defaults on those debts. Now, however, loans would be originated and subsequently be put through a process called ‘securitisation’ (Allen and Carletti, 2010), which meant that they would not be held on the banks’ balance sheets, allowing them thereby to bypass Basel’s capital requirements (Acharya and Richardson, 2009) and off-load risk (Brunnermeier, 2008). This, in turn, led to a shift in incentives as mortgage originators were more concerned with the number of mortgages they could initiate than with the debtors’ ability to repay them (Allen and Carletti, 2010). As house prices were going up at the time, even in the case of default, the collateral the bank would be able to repossess would have become more valuable (Ackermann, 2008).

There was an increase in subprime lending, where people with bad credit histories and low levels of income could be offered a mortgage. The number of new mortgages of this type originated grew from 6% to 15% between 2001 and 2006. Whilst standards fell further with NINJA loans being given out, referring to borrowers with no income, no job and no assets (Ackermann, 2008). The move towards the ‘originate and distribute’ model ultimately worsened the quality of mortgages that were originated (Purnanandam, 2011).

Securitisation involved pooling together a number of mortgages; this would create a product called a mortgage-backed security (MBS). In turn, this security would pay the holder a fixed income though the payments made towards the interest and principal of the underlying mortgages (Acharya and Richardson, 2009). Collateralised debt obligations (CDOs) were also used in the same manner as MBSs,
here the security consisted not only of mortgages but also other products, such as credit card debt and corporate bonds (Brunnermeier, 2008). These securities were then cut up into tranches of varying risk levels, where those deemed as having a high level of risk paid a higher premium compared to those with a lower level of risk (Acharya and Richardson, 2009). These MBSs were rated by the rating agencies (Patronoy, 2009).

1.1.2 Rating agencies

Rating agencies assign ratings to securities in relation to their probability of default (Cantor and Packer, 1994). This is done through factoring in various different forms of information regarding the issuer, the market, the specific security, and the macroeconomic climate (IOSCO, 2003). These ratings of credit risk alleviate the market from information asymmetry, helping investors by reducing uncertainty and allowing issuers to lower their cost of capital (IOSCO, 2003). Each company has its own rating system, but they are essentially harmonised on a notch basis. For Moody’s, these range from the highest rating AAA down to C, whilst for Standard & Poor’s and Fitch, they range from AAA to D. (Moody’s, 2015; Standard & Poor’s Rating Services, 2015; Fitch Ratings, 2014). The higher the rating the agency assigns, the lower the probability of default (Ferri et al., 1999) while a rating of BBB is considered the cut off in terms of investment grade securities (Standard & Poor’s Rating Services, 2015). Rating agencies are influential as governmental bodies such as the Securities and Exchange Commission (SEC) only allow money market funds to invest in securities rated AAA and AA (Bahena, 2010). Furthermore, if a security’s rating is upgraded, its price increases, whilst if it is downgraded, the price decreases (AAII, 2015).

There are only three major players in the rating industry, Moody’s, S&P and Fitch (Ferri et al., 1993) and they are dubbed the ‘Big Three’ (White, 2009). Indeed, 97% of the total market revenue goes to these three companies (SEC, 2012).

The recent history of rating agency performance has not been one to boast about. They were not able to foresee the East Asian crisis and in turn downgraded the countries involved by a more than proportionate amount. This had detrimental effects on their economies as it increased their cost of capital and lowered the amount of foreign capital flows into the respective countries (Ferri et al., 1999). The Big Three also gave an investment grade rating to Enron only four days before it went bankrupt (Hill, 2003; Hill 2009).
More recently, rating agencies have been condemned for giving inflated ratings to securities, which in turn spurred the market for MBSs and increased house prices (Patronoy 2009). Moody’s downgraded 54% of the subprime tranches it had rated in 2006 and 39% of those it had rated in 2007. S&P downgraded 44% of those originally rated between Q1 2005 and Q3 2007 and Fitch downgraded 34% for those in 2006 up to Q1 2007 (Hill, 2010). The number of these downgrades can therefore lead one to question the validity of the ratings the Big Three assign to securities. Indeed S&P have now admitted wrong doing and have reached an agreement with the Department of Justice to pay a $1.375bn fine for defrauding investors who were using their MBS and CDO ratings which were inflated and did not reflect the true credit risk of the underlying securities (Department of Justice, 2015).

1.2 Significance of the research
The significance of this research project is to fill the gap in the literature regarding people’s perceptions of credit rating agencies and their respective shortfalls. Furthermore, this research project attempts to link the existing theory surrounding the topic with academics and professionals’ views and looks for areas of agreement and disagreement.

1.3 Aims and objectives
1.3.1 Aims
- To understand the perceptions academics and professionals have of rating agencies and their shortfalls following the US financial crisis.

1.3.2 Objectives
- Conduct in-depth semi-structured interviews with a sample of academics and professionals with a background in finance or economics.
- Compare the data collected with existing literature to establish whether the sample’s perceptions are in accordance with the relevant theory.

1.4 Structure of the dissertation
This dissertation will be made up of 5 Sections, the first of which has provided the reader with a background to the rating agency industry and to structured debt in addition to having outlined the significance of the research along with the researcher’s aims and objectives. Section 2 provides a literature review to the topic of rating agencies. The section will be split up into four sub-sections: (a) The role of rating agencies in the financial crisis, (b) Changing business model and conflict of interest, (c) Complexity and modelling, and (d) Competition. Section 3 will discuss the methodology used to
conduct the study, the in-depth interviews, along with a discussion of the analysis used – thematic analysis. This will be followed by the ethical considerations that were kept in mind throughout the process of conducting the research project. Section 4 will analyse the data collected from the interviews and compare it with the existing literature on the topic of rating agencies. Finally, Section 5 will conclude the research project. It will provide a summary of the findings followed by the limitations of the research study and areas in which it can be developed.

2. Literature Review

2.1 Introduction

This Section contains a critical review of the relevant literature and empirical evidence in regard to the role that rating agencies played in the US financial crisis along with their shortfalls. Due to the scarce academic literature on the topic, newspaper articles and reports have been included as they provide valuable information towards the subject.

2.2 The role of rating agencies in the financial crisis and their reputations

Several academics are of the belief that rating agencies were at the centre of the financial crisis, as the inflated ratings they issued for MBSs spurred the market. This, in turn, led to more MBSs and similar financial products being demanded from the buy-side (e.g., Investment Funds) and consequently supplied by the sell-side (Investment banks), which exacerbated the total amount of losses to investors and banks when people defaulted on their debts (White, 2009; Patronoy, 2009; Patronoy 2010; Utzig, 2010). Others are of a similar view, but don’t consider rating agencies as the main culprits in causing the crisis as MBSs were mainly sold to banks, who would have had the skills necessary to determine the risk associated with these products, rather than individual investors. Thus, they were not marketing these securities to investors through inflated ratings (Acharya and Richardson, 2009). Hill (2010) underlines this point: “These instruments were not bought by widows and orphans” (p.598). However, though, Ackermann (2008) blames investors for not being thorough enough when making investment decisions, he does see a point for blaming the rating agencies for not giving ratings of a high quality. Rating agencies have since had their reputation tarnished (Bahena, 2010) leaving some to question the credibility of their ratings (Ryan, 2012). Indeed the CFA Institute (2014) provides evidence that investors have become more wary when looking at ratings for investment purposes.
2.3 Changing business model and conflicts of interest

As White (2009, 2010) recalls, the change in business model from an ‘investor pays’ principle to one where the issuers pay for a rating in the 1970s gave rise to opportunities for conflict of interest. Rating agencies would be able to guarantee future business by giving more favourable ratings. Although, this may not be applicable to companies wanting ratings for their corporate bonds due to the large number of them, there were only a few investment banks wishing to get large volumes of MBSs rated, creating an incentive to remain on favourable terms with these clients as losing a single one of them would already have resulted in a significant loss of business (Hunt, 2009; White, 2009; White 2010). Patronoy (2009) is also of the view that a change in the underlying business model had had a negative effect on the way rating agencies operated and believes that the loss in reputation from giving incorrect ratings would have been small relative to the profits they would have made. In addition, as the number of securities rated increased dramatically, the resources used per security-issue rating fell. Strobl and Xia (2012) support this argument with empirical evidence, using proxies for conflicts of interest, they find that Standard & Poor’s, which adopts an ‘issuer pays’ model, is more likely to give a higher rating than Egan-Jones Rating Company, which adopts an ‘investor pays’ business model. Hill (2010) does, however, bring up the point that rating agencies also rate issuers’ corporate bonds, which are rarely given the premium ‘AAA’ rating and that agencies would suffer a loss of reputation if they were to give unjust ratings, thus giving rise to the possibility of a loss of all future business. Whilst the European Commission (2008) refers to how over one tenth of 1,956 investment professionals had seen ratings change because of external pressures (European Commission, 2008). The reason for 51% these were to avoid losing business whilst 17% were to increase business (European Commission, 2008). While the majority of CFA Institute members believe that rating agencies still feel obliged to avoid downgrading or assigning higher ratings than deserved (CFA Institute, 2014). Indeed there are calls for a change in the current business model, although a switch back to the investor-pays model would rid the industry of the aforementioned conflicts of interest there is the ability to be a free-rider in the market. A government subsidy funded by issuers could in turn make sure that credit rating agencies have access to the appropriate resources (Deb and Murphy, 2009).

2.4 Complexity and modelling

Although structured products such as CDOs have differing factors of risk to corporate bonds, they were rated on the same ratings scale (Department of the Treasury 2008; Fender and Mitchell 2005, Hunt, 2009). This was done so that investors would be happy to buy the new complex securities (Fender and Mitchell 2008) without knowing the difference (Department of the Treasury 2008). Indeed, these structured products did not act in the same way as corporate bonds, as they were more
likely to get downgraded during the crisis, in turn, making people question whether they should follow the same rating scale (Hunt, 2009).

The securities that rating agencies were given to rate became more complex (SEC, 2008, Bahena, 2010) and as the complexity of an asset increases, models used to measure their underlying risk become less reliable (Danielsson, 2008). Fender and Mitchell (2005) add to this point in stating that due to added complexity through tranching and pooling, the ratings of CDOs would only give a partial view on the real underlying risks of the security involved. While Sketra and Veldkamp (2009) believe that as an asset becomes more complex there is more room for differing ratings. This creates the incentive for issuers to ratings shop and thus spurs ratings inflation even if they are unbiased.

Brunnermeier (2009) brings up the point that these models were too optimistic and did not count for a systematic fall in house prices to which Danielsson (2008) agrees and lists other modelling failures such as lack of proper documentation and screening. While Hunt (2009) summarises their modelling shortfalls as: not having realistic MBS default risks, lack of historical data to input into models and rating agencies not reacting to the lowering of mortgage standards. Some academics believe that CDO specialists did not even have the expertise to evaluate the risks involved with the assets that were susceptible to default correlation (Danielsson, 2008; Duffie, 2008; Hunt, 2009 Patronoy 2009). There was a lack of understanding of the risks involved with these sorts of securities (Teply et al., 2010).

Rating agencies may have also had a false sense of security in the build up to the crisis as they were ‘drinking the Kool-aid’, referring to how they were losing sight of the bigger picture (Hill, 2010). Indeed “the structured finance hype caused the [credit rating agencies] to lose their grip on reality” (Bahena, 2010, p.9). Although they are paid to assess the credit quality of securities, many other participants who felt the effects of the financial crisis did not understand the products or foresee the housing bubble bursting either (Hill, 2010). Several market participants lost track of value judgement and were lured by high ratings (Patronoy, 2009).

2.5 Competition

The lack of competition in the rating agency business has been highlighted in the literature (Economist, 2005; Hunt, 2009; Ryan 2012) and several believe that this is the reason for the low quality ratings issued by rating agencies (IOSCO, 2008). Hunt (2009) mentions, how in an environment with more competition, there would be more chance of a competitor producing high quality ratings. This lack of competition has been made about through the fact that only the SEC can decide who can provide
ratings for the creditworthiness of structured debt (Hunt, 2009; The Economist, 2005). Some see this as being a barrier to entry as if you were not given the ‘nationally recognised statistical rating organization’ (NRSRO) seal you would be side-lined from the ratings business (Hunt, 2009; White, 2009). Although by 2009, the number of NRSROs increased from 3 to 10, only the three biggest rating agencies had a real foothold in the industry, S&P, Moody’s and Fitch (White, 2009). In contrary, Hill (2010) believes it is the stickiness of investors which makes the industry so uncompetitive rather than the NRSRO seal, as there should have been a substantial loss of market share from the Big Three after Enron went bankrupt, however this did not happen.

Hill (2009) brings up the point that sometimes securities would need specific rating agencies to rate their security, so there would be no incentive to inflate ratings on these specific securities. Whereas Ryan (2012) is of the belief that there is a duopoly in the market if we disregard Fitch from the Big Three. Thus if 2 ratings are needed for a specific security, S&P and Moody’s would not need to compete, Hunt (2009) refers to this situation as a “partner monopoly” (p.132) amplifying this lack of competition.

Hunt (2009) also brings about the point of competitive laxity, where competition would actually lower the quality of ratings by inflating them in order to win clients. Indeed when Fitch entered the ratings market; there was a fall in the quality of ratings (Ryan, 2012). Becker and Milbourn (2011) were able to provide evidence for this by showing that a positive correlation existed between Fitch’s market share in the ratings industry and the grade of ratings produced, inferring that an increase in competition would yield higher ratings, inferring lower quality ratings. Furthermore, Strobl and Xia (2012) found that S&P would give higher ratings under a more competitive environment (if they transitioned from being the sole rater for an issuer to not being the sole rater).

While Jones (2008) refers to how Moody’s abandoned their diversity score initiative, which would yield lower ratings for securities with pools of the same collateral such as CDOs, and subsequently saw their business in this field prosper.

2.6 Summary

Many believe that rating agencies played a central role in the US financial crisis as their inflated ratings for structured debt instruments spurred the market, in turn amplifying the negative effects when the market was in a downturn. However, these structured debt instruments were not solely bought by retail investors, they were mainly bought by financial institutions who would have had knowledge of
the underlying risks these securities were subject to. Thus the rating agencies’ role in the lead up to the recent recession may be limited. With this said, their reputations are believed to have been tarnished following the financial crisis. A change in the underlying business model brought about a conflict of interest, where it would be in the raters’ interests to supply favourable ratings to guarantee future business. As securities became more complex, the methods used to determine their underlying risk became less effective, and the respective models came to rely on increasingly less realistic assumptions. Further, there was also a lack of competition in the industry that may have been a reason for the low quality of ratings. However, there is also evidence to suggest that an increase in competition brings forward the idea of ‘competitive laxity’.

3. Methodology

3.1 Introduction

The aim of this Section is to provide the information and reasoning behind the methods used to collect the data for this research project. The advantages and limits to semi-structured in-depth interviews and thematic analysis will be discussed, followed by the ethical issues that were kept in mind throughout the entirety of the project.

3.2 Qualitative research

This research project applies a qualitative approach. This method of research was used as it is an inquiry into the way people think and why. The samples are small and the data obtained refers to opinions as opposed to quantitative research, which uses large samples providing measurements (Keegan, 2009). This enables the researcher to answer the question proposed as it is not binary or something one can easily measure, instead the answer is more complex and based on people’s perceptions. Furthermore, since the pioneering work of Keynes (1936) the role which perceptions, conventions and expectations play in the financial markets has been largely acknowledged.

3.3 Method

The researcher gathered data through the use of semi-structured in-depth interviewing. A total of eight participants with backgrounds in economics or finance took part. Each one of these interviews lasted around 60 minutes and was either conducted on university campus or over the phone. Following the interview, the researcher transcribed the conversations and analysed them. Thematic analysis was used to analyse the data through coding the transcript to look for relevant themes in order to compare them with the literature that exists on the topic.
3.3.1 Semi-structured interviews
DiCicco-Bloom and Crabtree (2006) point out that semi-structured interviews are normally the sole research method used in order to gather information for a qualitative research project. Semi-structured interviews are a good way to gather intelligence from participants through the use of questions (Longhurst, 2010). They allow the researcher to gather comprehensive intelligence on people’s perceptions and on topical issues (Boyce and Neale, 2006). They can yield better information than other research methods such as surveys as the participant, being in an informal environment, may find it easier to share information with you (Boyce and Neale, 2006).

For these reasons, the researcher has decided to conduct their research in such a way; so that they may be able to directly understand the perceptions academics and professionals have in relation to rating agencies and their shortfalls following the US financial crisis.

3.3.2 The sample
Purposeful sampling can be used, where participants have a connection to the research question so that the information gathered from the interview process can be of a high level (DiCicco-Bloom and Crabtree, 2006). Therefore, the researcher has interviewed professionals in the financial services sector along with economic and financial academics. Table 1 illustrates the participants’ respective backgrounds, with the list ranging from an ex-hedge fund manager to economic research fellows.

When general themes start to emerge between participants’ views, the sample is considered to be of satisfactory size (Boyce & Neale, 2006). For this reason the researcher decided not to put an initial cap on the number of interviews conducted so that they would be able to explore the relevant themes surrounding the topic.
### Table 1: Participants’ background

<table>
<thead>
<tr>
<th>Participant #</th>
<th>Participant background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>Investment professional at a leading private equity firm</td>
</tr>
<tr>
<td>Participant 2</td>
<td>Director of sales at a boutique asset management firm</td>
</tr>
<tr>
<td>Participant 3</td>
<td>Principal teaching fellow in accounting and finance at a leading British university</td>
</tr>
<tr>
<td>Participant 4</td>
<td>Economic research fellow at a leading British university</td>
</tr>
<tr>
<td>Participant 5</td>
<td>Associate partner at a leading investment management firm</td>
</tr>
<tr>
<td>Participant 6</td>
<td>Partner at a leading investment management firm</td>
</tr>
<tr>
<td>Participant 7</td>
<td>Ex-hedge fund manager</td>
</tr>
<tr>
<td>Participant 8</td>
<td>Analyst at a leading financial advisory firm</td>
</tr>
</tbody>
</table>

#### 3.3.3 The questions

Open-ended questions are asked so that participants are able to share their views and experiences (DiCicco-Bloom and Crabtree, 2006). These questions will have been written up prior to the meeting and run through in a conversational setting during the interview (DiCicco-Bloom and Crabtree, 2006; Longhurst, 2010).

In order to assure that the relevant topics regarding the research project are all discussed, the researcher should make sure they consult with the participants beforehand and identify the questions that will be asked during the interview (Boyce and Neale, 2006). When arranging timings for the interviews the researcher was to conduct, they notified the participants of the questions they were looking to ask (Appendix A), so that they would be able to offer information on the context to the interview.
3.3.4 Limits to semi-structured interviews

Interviews are prone to biases (Boyce and Neale, 2006); so steps should be taken in order to keep those at a minimum. I have, therefore, avoided interviewing rating agency professionals to avoid the data collected to be skewed. With this said, it does not allow us to understand the research question from the ‘rating agency’s point of view’. Interviews are a long-winded research method as the researcher must dedicate time towards interviewing participants and recording the conversation, subsequently transcribing the recording (Appendix B), followed by analysing what was said (Appendix C) (Boyce and Neale, 2006). However, this time consuming process also allows the researcher to gain a better understanding of the data (Braun and Clarke, 2006).

There are other ways to collect qualitative data such as focus groups, where a cluster of participants share their thoughts on a specific topic; these lack, however, the depth of a one-to-one interview and are not structured in a way to gain information from several different participants in a shorter timeframe (DiCicco-Bloom and Crabtree, 2006).

3.3.5 Thematic analysis

Thematic analysis is a way the researcher is able to analyse data they have collected through discovering patterns within it. These patterns relate to groups of similar responses that are relevant to the research question (Joffe and Yardley, 2004; Braun and Clarke, 2006). The researcher reads through the data, in this case the transcripts from interviews and codes them, pointing out interesting snippets of data, which can relate to the research project. Once these transcripts have been analysed with different codes, the researcher then analyses the codes themselves in order to look for general themes among the data collected (Braun and Clarke, 2006). Themes that derive from theory allow the researcher to accept or contest the relevant literature (Braun and Clarke, 2006). Thematic analysis also provides a flexible way to analyse data as it is free from theoretical knowledge and can be used in conjunction with different sorts of theoretical beliefs (Braun and Clarke, 2006).

3.3.6 Limits to thematic analysis

Researches must be weary of invalidating the data as when thematic analysis is used, certain phrases may be taken out of context; also, the repeated presence of codes does not imply greater importance. For example, the repeated code of ‘pain’ may not imply a greater amount of pain being felt by an individual, but an increased will to talk of the subject (Joffe and Yardley, 2004).
3.5 Ethical issues

Interviewers should keep in mind what repercussions their actions may have as the interview process can impose psychological harm upon the interviewee. Furthermore, they must maintain the anonymity of the participants (Longhurst, 2010) to avoid the dangers that may arise if information of participants comes into the hands of those with differing interests (DiCicco-Bloom and Crabtree, 2006). For this reason, the researcher made sure the participants were conscious of the fact that their identities would be kept anonymous. The researcher should also provide ample information regarding the reason for interviewing the participants along with being given consent from the interviewees (DiCicco-Bloom and Crabtree, 2006). For this reason, the researcher sent an information sheet (Appendix D) and a consent form (Appendix E) to each participant prior to the interview.

3.6 Summary

The researcher has explained the reason for using an in-depth interview method along with thematic analysis in order to gather and analyse data needed in order to answer their research question. The researcher has also kept in mind several ethical considerations, to ensure that their actions have no negative consequences on other parties.

4. Analysis and Findings

4.1 Introduction

This Section will run through the analysed transcripts from the eight interviews that have taken place and compare them with the literature on and theoretical views about the role that rating agencies have played in the run-up to the financial crisis. Four general themes were quite evident, loss of reputation, conflicts of interest, complexity and lack of judgment, along with a general lack of competition in the ratings industry.

4.2 Loss of reputation for rating agencies

Whilst interviewing the participants, it was quite obvious the there was a theme of ‘loss of reputation’ for rating agencies following the US financial crisis. Indeed, seven out of the eight participants were in agreement to the statement in table 2.
Table 2: Sample’s perceptions of rating agency reputation following the U.S. financial crisis

| The US Financial crisis has negatively affected the reputation of rating agencies |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| Strongly Agree                  | Agree                          | Unsure                          | Disagree                        | Strongly Disagree                |
| 2                               | 5                              | 0                               | 1                               | 0                               |

Interviewee 2 saw their sluggishness in dealing with the problem as a reason for the loss of reputation:

“The rating agencies did not downgrade in advance. They were slow to react to the problem and therefore they have paid the price in terms of a loss of reputation.”

While interviewee 8 stressed that the financial sector has been damaged as a whole:

“The whole of the financial system, including the rating agencies, has been negatively affected.”

Others expressed how the ratings that are now assigned to securities lack credibility, and that they have changed their behaviour in terms of how they use ratings provided by credit rating agencies. For example, interviewee 2 said:

“I still think that people will view their ratings mildly sceptically.”

Interviewee 4 supported this view in saying:

“I am more suspicious now, they got it wrong and they still haven’t got the basics right.”

These views are in accordance to Bahena (2010) - reputations of rating agencies were negatively affected, questioning the credibility of their ratings as Ryan (2012) does. They also complement the CFA Institute’s findings where 67% of its members agreed that investors now look more vigilantly when using ratings do decipher potential investments. With this said, interviewees 5 and 8 still gave credit to the rating agencies, which was something that was not found in the academic literature.

Interviewee 5:

“Whilst rating agencies still perform a valid function there is now more emphasis on the need for our own research to complement their views.”
Interviewee 8:

“They’re still the best around, regardless of the past. They may have messed up, however, they still provide a valuable service to investors”.

Interestingly, interviewee 4, was in disagreement with the general rhetoric, and questioned whether the $1.375bn fine S&P had to pay (Department of Justice, 2015), changed anything:

“In theory I agree, their reputation has been negatively affected, but in reality, with the power they have, nothing has changed. Rules haven’t changed on their ratings and nobody has done anything about it. Although, yes, they paid $1.35bn or so, the failure of Lehman Brothers cost the taxpayer around $300bn. $1.35bn is peanuts, this is just PR money only to say that they’ve done something. I would rather that they didn’t pay, but had their [business] model restructured instead so that we can avoid something like this happening again in the future”.

All of the participants alluded to the view that rating agencies played a role in the financial crisis, however, not one suggested that they were one of its main causes; thus, disagreeing hereby with the view expressed in the works of White (2009), Patronoy, (2009, 2010) and Utzig (2010), who consider inflated ratings a central cog in the works of the financial crisis, stimulating the market for MBSs. Interviewee 7 went on to suggest how they worsened the impact of the financial crisis without having been, however, its initiators.

“Investors relied on rating agencies more than they should have, however I don’t think it was inflated ratings that caused the financial crisis, although it allowed pension funds etc. to get into the market. The principal players were the leveraged buyers”.

4.3 Conflicts of interest

When the participants were questioned about whether a shift in the underlying business model would change the way credit rating agencies operate, the data obtained was inconclusive, with four agreeing with the statement, one disagreeing and three unsure (see Appendix F). However, all participants expressed doubts over the efficiency of an “issuer pays” business model throughout the interview process and a theme of conflicting interests developed. Interviewee 1 shared the view of the European Commission (2008) in thinking that the current “issuer pays” model leaves room for issuers to push for higher ratings:
“In the end of the day there will always be pressures from companies to get higher ratings”.

Interviewee 2 agreed:

“I feel that if issuers are paying, these companies will be putting pressure on the rating agencies in order to get more favourable ratings”.

So did Interviewee 4:

“Generally, changing the stakeholders shifts the level of interest towards them. Companies will defend their interests and exercise the power they have in the rating process”.

Some even put forward the idea of moving back toward an “investor pays” business model, in hope of ridding the industry of conflicting interests echoing the views of Deb and Murphy (2009) and the majority (52%) of CFA Institute members (CFA Institute, 2014). Interviewee 7 said:

“The rating agencies played both the gamekeeper and the poacher. Investors needed to be protected by the rating agencies, how can they be protected if their interests aren’t aligned? They should be paid by investors rather than issuers. They are half-way between civil servants and bankers, they were enticed by the higher levels of compensation they would get from rating these exotic instruments which would make them want to keep the investment banks happy by giving better ratings. The industry should find a way of returning to an ‘investor pays’ business model”.

However, some of participants pointed out the difficulty in adopting a new “investor pays” model, as it gives room for free riders to enter the market. Agreeing with one of White’s (2009) reasons for which the credit rating agencies’ business model changed in the first place, as due to technological developments, ratings can freely be shared among investors. For example, interviewee 3 said:

“But, how can you pay for a public good, it would be difficult to be paid by investors”.

Interviewee 8 was in agreement:

“There should be a move back to the ‘investor pays’ model, however, the same problems of freeriding would come about, just as they did in pre-1970s”.
One participant however, interviewee 6, saw this conflict of interest as being less important in the run up to the financial crisis, although believing that there is a problem with the current business model, they placed more emphasis on the lack of expertise on of the raters:

“The rating agencies failed to identify the risks being taken and gave AAA ratings to CDOs, which turned out to be junk. I would question the relationship between the agencies and the issuers of debt. [...] The move from a subscriber (investor) to an ‘issuer pays’ model occurred in the 1970s, a long time before the recent US financial crisis. Their behaviour did not change, but their analysis of CDS did not work”.

4.4 Complexity and lack of judgment

Overall, the participants saw the complexity of CDOs and the models used to measure their credit risk as being problematic. Indeed, three-quarters agreed to the statement in Table 3. This idea of complexity was a prevalent theme throughout the interviews that were conducted, whilst the theme of a lack of judgment also became clear.

Table 3: Sample’s perceptions on whether there was a problem with the complexity of the securities and the inadequate models used to measure risk

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Unsure</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>There was a problem with the complexity of the securities and the inadequate models used by rating agencies to measure risk</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In accordance with Danielsson (2008), interviewee 1 allured to the idea that there was amplified difficulty in determining the risk associated with these securities as their complexities increased:

“CDOs are very complex, the more complex the product the more complex the modelling. I can imagine they were very difficult to understand, and thus with more complexity the less informed you would be. It is difficult to understand the underlying risk if you do not understand the product”.

Interviewee 4 agreed with this idea, and in support of Danielsson (2008) Duffie (2008), Hunt (2009) and Patronoy (2009), believed that rating agencies lacked the expertise needed:
‘The more unknowns you bundle together, the more chance that something will go wrong, this is what was happening and it created a huge lack of information. People didn’t have the knowledge to understand the underlying risks in CDOs’.

While Interviewee 8 referred to the lack of understanding that was inherent throughout the financial services industry, supporting Teply et al. (2010).

“The securities became so complex that nobody really knew what was in them. This created a flawed process as how can you rate something you don’t understand”.

A theme that stemmed off this idea of complexity was lack of judgement. Several interviewees took this point of view as they thought that market participants and rating agencies were lured by the profitable CDO business without questioning its long-term reliability and thus ‘drinking the Kool-aid’ as referred to by Hill (2011).

For example, interviewee 2 and 4 stated they were fuelled by greed and were not using common sense.

Interviewee 2:

“The real problem was that people lost the sight of common sense when it came to the complexity of the securities. You could compare this to the story of Icarus; people were seeking higher and higher returns and just lost sight of common sense until everything came spiralling downwards”.

Interviewee 4:

“They were enjoying the fruits of a booming market with a ‘don’t mess with success’ attitude and so were blindly continuing with what they were doing as they were making money and it seemed as if it was working”.

While interviewee 1 and 3 said that they couldn’t see the inherent risks associated with their actions.

Interviewee 1:

“They were all working in a cloud where they couldn’t see what was actually happening.”
Interviewee 3:

“As you get into more and more complex markets, you lose the idea of the big picture in terms of calculations. You could get blinded by the models, without thinking of other factors within the market. Complexity is a given, however, basic judgement still needs to be there. Everybody was caught up with the magic that was going on; banks, investors, regulators and rating agencies etc.”

Although interviewee 7 thought the rating agencies used assumptions that were too optimistic in their models, in line with Brunnermeier (2009), they placed more emphasis on investor behaviour:

“At issuance, the ratings and prices may have been more or less correct, but as the crisis unfolded, people who owned these securities didn’t want them anymore, and sold them whether or not the price was in line with the probability of those mortgages defaulting. In fact, most of those mortgages didn’t even default; it was just the behaviour of investors that made the price fall to such a level. The government actually managed to make quite a lot of money when they bought these ‘toxic assets’ from banks and held them until maturity!”

4.5 Competition

Overall, the majority of participants believed that the level of competition in the industry affected the way they operate, as the data shows in in table 4.

Table 4: Sample’s perceptions on whether competition affects ratings

<table>
<thead>
<tr>
<th>The level of competition within the rating agency industry affects ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

All but one of the participants agreed with the idea that the rating agency industry was uncompetitive (Economist, 2005; Hunt, 2009; Ryan 2012), for example interviewee 4 stated:

“[…] they ended up having too much power, the barriers to entry in the market were so high that there was an effective oligopoly”. 
Interviewee 3 commented:

“There are only about 3 or 4 rating agencies that have a massive share of the market”.

Whilst interviewee 2 echoed the view of the IOSCO (2008) in thinking that the lack of competition could lead to unwarranted ratings. They also brought up the fact that there are measures put in place to limit this behaviour, as referred to by Hill (2010):

“If there is an oligopoly, as is the case for Moody’s, S&P and Fitch, there is potential for the output to be skewed. This manifested itself in terms of higher ratings. However, there is still need for 2 ratings for certain securities”.

While interviewee 1 also brought up the need for 2 ratings, they made the point that these ratings are mostly the same, enhancing this lack of competition as Hunt (2009) did when describing the two biggest rating agencies, S&P and Moody’s as having a “partner monopoly” (Hunt, 2009, p.132).

Interviewee 1:

“There are normally 2 ratings provided for a debt instrument precisely for this reason and there aren’t many split (differing) ratings”.

Several of the participants saw the need for more competition in the industry in order to increase the quality of ratings, agreeing with the theory put forward by Hunt (2009) where increased competition would lead to increased probability of higher quality ratings. For example Interviewee 5 said:

“Competition should have a positive effect on ratings as the quality of their judgments will be scrutinised in a greater fashion driving up performance. Any increase in competition should ultimately drive up the quality of the industry – which is the ultimate aim”.

Interviewee 4 was of a similar view:

“As regulators did not understand what they were doing when rating these CDOs they had no knowledge of whether they were acting efficiently or not. There was a need for other competitors who could figure this out and stop what was happening”.

Interviewee 7, however, was unsure on whether increasing competition would have a positive effect, touching on the theory of competitive laxity (Hunt, 2009):
“Competition did not make a difference with plain vanilla (basic) securities, when they became more complex it made them lower their standards. I’m not sure whether increased or decreased competition would have made the situation better or not”.

Whilst Interviewee 8 was in agreement with Hill (2010) in thinking that the uncompetitive nature of the industry came about through investor stickiness rather than a lack of competitors:

“The lack of competition means that the established firms, e.g. Moody’s, are trusted, people will carry on using them”.

5. Conclusion

The final Section of this dissertation contains a summary of the research findings from the interviewing process. It will also cover the limitations to the research conducted and discuss avenues in which further research can develop and build upon this research project.

5.1 Summary of findings

The researcher has kept in line with his aims and objectives by conducting in-depth semi-structured interviews with eight participants, all of whom had backgrounds in economics or finance. In doing so, he was able to understand their perceptions regarding rating agencies and their shortfalls and compare them with the existing literature surrounding the topic.

Indeed, all but one of the participants shared the view that there has been a loss in reputation for rating agencies following the US financial crisis. With all of the interviewees believing that rating agencies played a role in the financial crisis, some have started to question the validity of the ratings they produce and use them in a more cautious manner following the US financial crisis. Interestingly a couple of participants shared the view that rating agencies are still a valuable source of information regardless of their past performance.

When asked whether a change in their business model could have affected the ratings given to securities, the participants had differing views, thus making it difficult to come to a final conclusion whether it was a significant failure. However, as a whole they identified the failures inherent with this business model and the potential of conflicts of interest. Some participants longed for a shift back to the “investor pays” model, but saw difficulty in adopting it.
The majority of participants saw the complexity and inadequate models used to decipher the risk of MBSs and CDOs as one of the rating agencies’ major shortfalls. Several also believed that there was a lack of proper judgement in the rating agency industry where people stopped using common sense.

Finally, when asked whether one of their shortfalls could have been a general lack of competition within the industry, the majority agreed. With this said, three of the eight participants disagreed. The general view that an increase in competition within the market was necessary became evident whilst one participant explained that the uncompetitive nature of the market came about through investor stickiness.

5.2 Limitations of research

Only a small sample was used when collecting the data for this research project, which means that the sample’s views may not adequately reflect those of the population. Furthermore, although rating agency professionals were deliberately excluded from the sample in order to eliminate the risk of the bias that their inclusion would inevitably had introduced to the study, this leaves a gap in the research, as their views have not been considered in the researcher’s attempt to answer the question.

5.3 Further research

The research can be developed through using a larger sample of participants, so that stronger conclusions can be made surrounding the research question. Furthermore, it would allow the researcher to understand different groups of peoples’ perceptions concerning rating agencies’ reputations and their shortfalls and compare them with one another.

This research project can act as a building block for further research regarding perceptions of academics and professionals in regards to rating agencies’ reputation and shortfalls following the US financial crisis. Furthermore it could work towards a larger research project concerning peoples’ perceptions in relation to all the different driving forces of the financial crisis.
Bibliography


Appendices

Appendix A – Interview Questions

Interview Questions

1) In order of importance, what were the main causes of the US financial crisis and why?

2) The US financial crisis has negatively affected the reputation of rating agencies.
   
   A – I strongly agree
   B – I agree
   C – I disagree
   D – I strongly disagree
   E – I don’t know

   Why?

3) Has your behaviour changed in terms of how you use ratings provided by rating agencies following the US financial crisis?

4) Moving from an ‘investor pays’ business model to an ‘issuer pays’ business model has affected the way Rating Agencies operate.
   
   A – I strongly agree
   B – I agree
   C – I disagree
   D – I strongly disagree
   E – I don’t know

   Why?

5) The level of competition within the rating agency industry affects ratings.
   
   A – I strongly agree
   B – I agree
   C – I disagree
   D – I strongly disagree
   E – I don’t know

   Why?

6) There was problem with the complex models rating agencies used to measure risk.
   
   A – I strongly agree
   B – I agree
   C – I disagree
   D – I strongly disagree
   E – I don’t know

   Why?
7) Did rating agencies have any other shortfalls?

8) In what way could the rating agencies be reformed for the better?

Appendix B – Interview transcript example

Participant 1

1) That’s a hard question. I was not working at the time so I can only provide a reflection of what I’ve learnt through the years. The main cause of the financial crisis was Freddie Mac and Fannie Mae subsidising loans to those who couldn’t afford them, spurring investment in the housing market. Of course there were many other steps but the root of the problem was when Bush decided that every American should be able to have a house. This led to a credit bubble not only for people buying houses but also financial institutions. These institutions did not know what CDOs essentially were and what exactly were in them. It is clear that rating agencies were also rating these wrongly. S&P actually settled today admitting that they were mis-rating mortgages to get market share.

2) Agree, again there are wider concepts here. They were all working in a cloud where they couldn’t see what was actually happening. Yes it has, but not as badly as banks. There was more scrutiny on rating agencies being too aggressive with ratings at the impact of the crisis. We should assume that Moody’s would have to deal with a settlement as S&P did as they were the first non-bank to assume responsibility.

3) Don’t know.

4) Don’t know. In the end of the day there will always be pressures from companies to get higher ratings

5) Agree. There are normally 2 ratings provided for a debt instrument precisely for this reason and there aren’t many split ratings.

6) Agree. CDOs are very complex, the more complex the product the more complex the modelling. I can imagine they were very difficult to understand, and thus with more complexity the less informed you would be. It is difficult to understand the underlying risk if you do not understand the product.

7) Sorry, can’t help you with that.

8) The solution would be standardised in an ideal world were guidelines would be put in place. However, in reality, because of financing, companies are so financially complex and do not have the same template. It wouldn’t work to have guidelines as ratings could end up being misleading. Furthermore, if you can’t have a different methodology when rating, you cannot differentiate and do good analysis. You could maybe make them accountable for their ratings but you can’t really as ratings are a tool to assist investors. They are free to choose whether they use this tool or not, whilst others can asses the risk themselves, in-house.
Appendix C – Thematic analysis example

The theme of ‘complexity’ is highlighted in green with ‘lack of judgement’ in yellow.

Interviewee 3:

Derivatives are complicated so the models assessing them would be complex. The judgement part of the assessment is not that obvious. As you get into more and more complex markets, you lose the idea of the big picture in terms of calculation. You could get blinded by the models without thinking of other factors within the market. Complexity is a given, however basic judgement still needs to be there. Everybody was caught up with the magic that was going on; banks, investors, regulators and rating agencies etc.

Appendix D – Information Sheet

What was the involvement of rating agencies in the US financial crisis? What were their main shortfalls? A commentary from economic and financial academics and members of the financial services industry

Invitation

You have been invited to take part in a research project. Before you make your decision of whether to take part or not, it is essential for you to understand the reason for which this research is taking place and what will be asked from you. Please read the following document with care and feel free to contact me if anything is unclear or would like additional information.

What is the purpose of the project?

The purpose of this research project is to understand academics’ and professionals’ perceptions of rating agencies and their shortfalls following the financial crisis.

Why have I been chosen?

You have been chosen to be a participant in this research project as you are an economic/financial academic or work in the financial services sector and thus have ample and substantial knowledge in the area of this research project.

Do I have to take part?

You are free to decide whether to take part in this research project or not. If you decide to be a participant in this research project you will be provided with a consent form to sign. If you decide to no longer take part in the research project you may withdraw at any time without having to state a reason.

What do I have to do?

You will be taking part in an in-depth, semi-structured interview over the telephone or in person. This should last about an hour. You will be asked up to 10 open-ended questions.
What are the possible disadvantages and risks of taking part?

This study may lead to the discussion of sensitive material in regards to your views on the rating agency industry, which may lead you to feel uncomfortable.

What are the possible benefits of taking part?

There are no benefits for people participating in this research project, however, hopefully this work will provide an informative perception of the rating agency industry.

Will my taking part in this project be kept confidential? What will happen to the results of the research project?

All the data that will be collected in this research project will be confidential. Your identity will remain anonymous. It is not probable that this research project will be published, if however it were to be, all the details of where and when this project can be accessed will be sent to you.

Will I be recorded, and how will the recorded media be used?

The audio recordings from the interview will only be used for analysis. They will not be used for any other reason without your written permission. Furthermore, no external parties will have access to the audio recordings for this research project.

Contact for further information

If you have any queries or would like any further information, please do not hesitate to get in contact. My e-mail address is: bn12tojs@leeds.ac.uk

Thank you for taking the time to read this information with care.

Kind regards,

Timothy Sapsford
Appendix E – Consent form

Declaration of Consent

I ....................................................... (print name) give consent to participate in data collection as part of Timothy Sapsford’s undergraduate dissertation research at the University of Leeds. The research is focused on academics’ and professionals’ perceptions of rating agencies and their shortfalls after the US financial crisis.

I understand the information provided will be:

- used sensitively and confidentially
- anonymised such that respondents cannot be identified from the final report
- stored in a secure location:
  - digital data will be stored initially on a password protected USB, and then transferred to secure folder on the university network
  - hard-copy data will be stored in a locked filing cabinet on University premises

- destroyed after the research is complete.

Participant:

Signed ...........................................................
Dated ...........................................................

Researcher:

Signed ...........................................................
Dated ...........................................................

Research Supervisor:

Marco Veronese Passarella
Leeds University Business School
Maurice Keyworth Building
The University of Leeds
Leeds
LS2 9JT
M.Passarellaleeds.ac.uk
Appendix F – The participants’ perceptions regarding whether a change in rating agencies’ business model affected their operations

Moving from an ‘investor pays’ business model to an ‘issuer pays’ business model has affected the way rating agencies operate

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Unsure</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
An econometric study into the impact of the current account on house price growth in the United States

George Vernon

This dissertation investigates the impact of the current account on house price growth in the United States between Q4 1977 and Q4 2013. An econometric analysis is used to estimate the effect of changes in the current account balance on house price growth over the period, whilst accounting for a range of control variables. The regression results show that throughout the period, negative changes in the current account balance were associated with higher levels of house price growth. This finding holds even when controlling for potential reverse causality. However, the econometric model fails to account for all of the variation in house price growth and as such omitted variable bias cannot be ruled out.
1. Introduction

This dissertation investigates the impact of the current account on house price growth. It is imperative that economists and policymakers understand the housing market. Not only does housing play an important role in society, providing shelter for citizens, but the market for it is linked with the stability of the broader economy (Hirata et al., 2013). A country’s current account records economic transactions between residents and non-residents involving goods, services and income, the balance of this account is equal to the difference between saving and investment for an economy (IMF, 2009). The effect of changes in cross-border flows on asset prices is uncertain.

Whilst there is agreement regarding the core determinants of house prices there is ambiguity regarding the role of the current account. Some researchers argue changes in the current account affect house prices, specifically that current account deficits are associated with higher house price growth (Bernanke, 2005, 2010; Aizenman and Jinjarak, 2009; and Justiniano et al., 2014). However, others argue changes in house prices affect the current account (Fratzscher et al., 2010; Laibson and Mollerstrom, 2010; Geerolf and Grjebine, 2014). Much of the existing literature in this area has deficiencies, with studies often being based on limited samples using correlation analysis or unconditional regressions. Researchers on both sides of the debate fail to account for the possibility that the other side’s position is correct and hence fail to dismiss it.

It is the deficiencies in the existing literature that my research seeks to overcome. I aim to conduct a robust empirical investigation into the impact of the current account on house prices. To do this I will use an econometric methodology, using data from the United States (US) in the period from Q4 1977 to Q4 2013. This avoids small sample bias and the difficulties faced when trying to draw policy relevant conclusions from cross-country studies in the presence of heterogeneity. My aim is to generate robust results and I will conduct my analysis in a way that corrects for potential reverse causality.

My results show that negative changes in the current account balance were associated with higher rates of house price growth throughout the sample period. This relationship holds even when controlling for changes in real GDP per capita, inflation, population, interest rates and the state of the economy. The use of control variables adds an element of robustness absent from existing literature. These findings also hold when controlling for potential reverse causality, through the use of lagged independent variables and instrumental variable regression. The robustness of my methodology adds to the reliability of my conclusions.
This dissertation is structured as follows. Section 2 begins by discussing the importance of the housing market. I then analyse existing literature on the determinants of house prices and finish by discussing each side of the debate regarding the role of the current account. I use this discussion to shape my research aims, focusing on addressing areas of weakness in the literature.

Section 3 outlines the methodology and data I will use to conduct my analysis. I begin by outlining the econometric models I will run and a test of hypothesis that will be important in determining the robustness of my results. I then explain my choice of variables and sample, before describing data on house price growth and the current account for the US.

Section 4 presents the results of my analysis. Here I discuss these and conduct exercises to test for robustness. I do this by testing hypotheses, constructing confidence intervals and running a sensitivity analysis. I further address potential reverse causality by using instrumental variable regressions. I finish by considering the limitations of my analysis and suggest how these could be addressed by future research.

Section 5 concludes. Here I revisit the motivations for my analysis and review what I have found. I focus on assessing whether or not I have achieved my stated aims and the value of my contribution.

2. Literature Review

2.1 Introduction

This Chapter reviews existing literature regarding house prices. Section 2.2 outlines the importance of the housing market. Not only does housing have an important societal function, but changes in this market affect the wider economy. Section 2.3 examines the existing literature on the determinants of house prices. Section 2.4 discusses research that specifically links house prices to the current account, examining the two perspectives on this relationship. In Section 2.5 I summarise the discussions.

2.2 Importance of House Prices

Hirata et al. (2013) list three reasons why understanding the housing market is important: (1) housing’s function as shelter, (2) the significant proportion of GDP accounted for by housing related expenditure, and (3), the fact that housing is the primary asset, and mortgage debt the primary liability for households in developed economies. The combination of these facts mean house price volatility can have significant effects on the broader economy (Hirata et al., 2013).
Tsatsoronis and Zhu (2004) find a positive link between house prices and credit growth; implying a risk of dual imbalances in the housing market and the financial sector. The housing market was at the centre of the 2008 financial crisis, which was caused by the combination of a house price and credit boom (Acharya and Richardson, 2009). Falling house prices caused a systemic collapse in the value of financial assets leading to financial market instability (Allen and Carletti, 2010). The presence of negative externalities originating from the housing market and effecting the financial sector make it imperative to understand the former in order to understand the risks to stability in the latter.

Finally, Barack Obama has described rising inequality and low levels of social mobility as “the defining challenge of our time” (Obama, 2013). Recently, the issue of inequality has engendered much debate amongst economists and politicians, especially since the publication of Thomas Piketty’s ‘Capital in the Twenty-First Century’ (The Economist, 2015). Rognlie (2015) finds that much of the change in national income shares in the US can be explained by changes in house prices. This suggests that it is crucial that those concerned with the distribution of income understand the housing market.

2.3 Determinants of House Prices

The market for housing works like a normal commodity market where the quantity exchanged and the price level are determined by supply and demand (Igan and Loungani, 2012). In the short run, the supply of housing is inelastic, due to a number of factors, including the time needed for new housing to be built. This means prices are driven by changes in demand (Algieri, 2013). My investigation will focus on factors that drive the price of housing in the short run.

Tsatsoronis and Zhu (2004) find changes in income, demographics and taxation affect the demand for housing and subsequently prices. Also important is the cost and availability of credit. Low interest rates increase the demand for housing and, ceteris paribus, the price of it by relaxing household borrowing constraints (Tsatsoronis and Zhu, 2004). There is also a positive link between house prices and inflation, as housing investments can be used as an inflation hedge (Tsatsoronis and Zhu, 2004). However, I would question the significance of this factor given how costly housing related investments are and subsequently their accessibility to the average citizen.

Hirata et al. (2013) analyse correlations between house prices and a range of economic variables, finding that prices are pro-cyclical. This suggests prices are positively related to changes in income and that the economic cycle is important. Correlation does not mean causation; however, these findings have been supported by other research. Igan and Loungani (2012) find house prices are driven by
changes in both income and population. They suggest the link between house prices and interest rates is weaker than that suggested by other authors, the extent of any effect being dependent on local market characteristics (Igan and Loungani, 2012).

There is general agreement regarding the core determinants of house prices in the short run. These can be listed as income, inflation, demographics, interest rates and the state of the economy. However, there is less consensus when it comes to the relative importance and significance of these determinants (Algieri, 2013).

2.4 Open Economy Considerations
The following sections discuss the role of the current account, with regard to house prices. Section 2.4.1 outlines research that argues current accounts affect house prices. Section 2.4.2 examines the counterview which suggests causality runs the other way.

2.4.1 Current Accounts to House Prices
A seminal contribution in this area is Bernanke’s ‘Global Savings Glut Hypothesis’. This begins by outlining an accounting identity, current account imbalances in each period, must be offset by an equal quantity of capital flows (Bernanke, 2005). A country running a current account deficit will have inflows equal to that deficit; a country running a current account surplus will have outflows equal to that surplus. The US current account deficit and the capital inflows that accompany it has contributed to strong house price appreciation, indirectly through downward pressure on interest rates (Bernanke, 2005). This phenomenon is not unique to the US. Bernanke (2005) notes that other countries including France, Spain and the UK also saw their current account balances move towards deficit in the period studied, whilst Germany and Japan moved towards surplus. The deficit countries experienced house price appreciation, the surplus countries did not (Bernanke, 2005). This suggests differences in current account positions lead to differences in housing market outcomes.

However, this finding is based on little empirical evidence and the observed correlation does not necessarily mean causation. It is based on observations of the period from 1996 to 2003. It is questionable whether the relationship holds when a longer time period is considered. To further investigate the link between the two variables a more robust approach is needed, focused on a larger sample.
In later work, Bernanke (2010) presented further cross-country evidence, showing that differences in capital flows, emanating from differences in current account positions, explain differences in house price dynamics. Plotting real house price growth against the change in the current account over a 5 year period for 20 countries shows a strong negative relationship\(^1\) (Bernanke, 2010). A shift towards a current account deficit is linked to higher house price growth. A regression was run resulting in an \(R^2\) of 0.31, suggesting 31% of the variance in house price growth can be accounted for by changes in the current account (Bernanke, 2010).

However, causation cannot be inferred because the regression appears to be unconditional and may suffer from omitted variable bias. This is apparent by considering the proportion of house price variance not accounted for by changes in the current account. There is not enough methodological information to convince me that this approach is robust. Furthermore, whilst regression results for cross-country studies are correct on average they may not apply to individual countries, making it difficult to draw policy relevant conclusions from them. Ferrero (2012) replicated this analysis, expanding the sample to 32 countries. The results were similar, a strong negative relationship between the two variables represented by a correlation coefficient of -0.64 (Ferrero, 2012). This supports the hypothesis that there is a negative relationship between changes in the current account and house price growth. However, correlation is not enough to infer causation and this study does not advance the understanding of the relationship between the two variables.

Aizenman and Jinjarak (2009) conducted an econometric study into the impact of the current account on house prices; covering 43 countries over the period from 1990 and 2005. In terms of empirical robustness, this study is preferred to the studies discussed above. They found lagged current account deficits are associated with higher house price growth even when controlling for variables including the interest rate and GDP (Aizenman and Jinjarak, 2009). This research also provides an insight into the transmission mechanism for these effects. Two channels are identified: (1) reducing interest rates by increasing the amount of saving, and (2), direct housing purchases by foreign capital (Aizenman and Jinjarak, 2009). The use of lagged values of the current account goes some way towards addressing the possibility of reverse causality. However, the conclusions of this study may not apply equally to every country in the sample making it difficult to incorporate this research into policy. Furthermore, the 15 year period studied is not long enough to draw reliable conclusions about this relationship.

\(^1\) See Slide 10 in Bernanke (2010)
Asici and Hepsen (2013) also conducted an econometric study into the impact of the current account on house prices. This was focused on Turkey between 2007 and 2012. They found a positive relationship between current account deficits and house price growth (Asici and Hepsen, 2013). Whilst this methodology adds empirical robustness, it has only been conducted over a short time period. Results may be affected by small sample bias. Furthermore, these results lack external validity in terms of their application to developed economies.

Justiniano et al. (2014) examined the effect of changes in capital flows on US house prices, finding that these were responsible for 25-33% of the growth in real house prices over the course of the 2000’s. Changes in capital flows are equivalent to changes in the current account (Bernanke, 2005). This is a similar proportion of the variance in price growth attributed to changes in the current account, as suggested by Bernanke (2010). The transmission mechanism for these effects is downward pressure on interest rates (Justiniano et al., 2014). Whilst this adds empirical rigour to the literature, the research is only conducted for a short time period, making no attempt to assess the link between the two variables before 2000. Therefore, it could suffer from small sample bias.

2.4.2 House Prices to Current Accounts

Fratzscher et al. (2010) analysed the impact of asset prices on the US current account, finding equity and house price shocks explain a third of the change in the trade balance over a 5-year period. Asset price appreciation increases household spending and business investment inducing a decline in the trade balance (Fratzscher et al., 2010). However, it is unclear why these changes must induce such a decline. The authors fail to show why an increase in consumption or investment would not be focussed domestically. The effect is an empirical matter and the authors fail to support their argument with evidence. Furthermore, they consistently fail to distinguish between the trade balance and the current account. The results above relate to the trade balance but there is no mention of results once the total current account is considered.

Laibson and Mollerstrom (2010) found changes in house prices explained 50% of the variation in current account deficits between 1996 and 2006 for 19 countries. However, their analysis fails to deal with the possibility of reverse causality and as such this cannot be ruled out. The failure to empirically prove their case is admitted by the authors who state “these correlations do not settle the issue of causation” (Basco, 2009; cited in Laibson and Mollerstrom, 2010, p. 371). Furthermore, as stated previously, findings from cross-country studies may be correct on average but may not apply to all countries in the sample.
Geerolf and Grjebine (2013) conducted an econometric study, using property taxes as an instrumental variable for house prices, for 40 countries between 1970 and 2010. They found house prices are an important determinant of current accounts, a 10% increase in prices leading to a decline in the current account balance of 1.7% of GDP (Geerolf and Grjebine, 2013). The authors use instrumental variables to address reverse causality, arguing property taxes are independent of macroeconomic factors that affect the current account (Geerolf and Grjebine, 2013). However, they fail to define what is meant by macroeconomic conditions. If one considers macroeconomic conditions to include domestic economic growth, then fiscal policy is unlikely to be independent of such factors. Factors that affect the current account may also affect property taxes and reverse causality could still be a problem. Findings may be biased and therefore unreliable.

2.5 Summary

In this Chapter I outlined why it is important to understand the housing market. I have shown that there is a consensus in the literature regarding the core determinants of house prices. However, there is disagreement regarding the role of the current account.

Existing research that examines the impact of the current account on house prices has three main weaknesses. Firstly, a lack of empirical rigour. Much of the analysis in this area relies on correlations and unconditional regressions, making it difficult to infer causation and draw reliable conclusions. Secondly, other than Aizenman and Jinjarak (2009), there is a failure to address potential reverse causality. As such this possibility cannot be dismissed. This is also the main criticism of literature that examines the impact of house prices on current accounts. Finally, there is an absence of research that places the most recent data in a historical context, with much of the focus on short sample periods. As such findings may suffer from small sample bias.

My research aims to add value to the existing literature by analysing the impact of the current account on house prices, using an econometric methodology to obtain robust results. This will subsequently lead to more reliable conclusions and I will be in a better position to infer causation. Furthermore, by accepting the possibility of reverse causality and conducting my analysis in a way that addresses it. My analysis will also utilise the most recent data available to generate policy relevant results. This will be combined with historical data to construct a sample that avoids small sample bias.
3. Methodology and Data

3.1 Introduction

The aim of this dissertation is to analyse the impact of the current account on house price growth. To do this I will use an Ordinary Least Squares (OLS) regression model. This chapter outlines the models and data I will use to conduct my analysis. Section 3.2 specifies the econometric models I will run. I am using multiple models to minimise the impact of omitted variable bias and address potential reverse causality.

Section 3.3 outlines the hypothesis test that will be used to assess whether my estimated coefficients are statistically significant and support the hypothesis that there is a relationship between the two variables.

Section 3.4 describes the key variables in the model, house price growth and the current account. Here I also explain my set of control variables, what they are and the theoretical basis of their inclusion.

Section 3.5 outlines the data sample I will be using in my model and Section 3.6 presents this data.

3.2 Econometric Models

Below are the four OLS models that I will use in the course of my analysis:

\[ HPG_t = b_0 + b_1 CA_t + u_t \]  \hspace{1cm} (1)

\[ HPG_t = b_0 + b_1 CA_t + \gamma Z_t + u_t \]  \hspace{1cm} (2)

\[ HPG_t = b_0 + b_1 CA_{t-1} + u_t \]  \hspace{1cm} (3)

\[ HPG_t = b_0 + b_1 CA_{t-1} + \gamma Z_{t-1} + u_t \]  \hspace{1cm} (4)

Where:

- \( HPG_t \) denotes house price growth
- \( CA_t \) denotes current account
- \( Z_t \) denotes a set of control variables
- \( u_t \) denotes an error term
- Sub-script \( t \) denotes period \( t \)
Model (1) tests the effect of the current account on house price growth. In all models, $HPG$ is the dependent variable and $CA$ is the independent variable. The coefficient $b_1$ gives the change in house price growth given a one unit change in the current account. Model (1) may suffer from omitted variable bias which if unaddressed could result in biased coefficient estimates and inaccurate conclusions (Stock and Watson, 2012).

To reduce this potential distortion, Model (2) incorporates the core house price determinants identified in Section 2.3 as control variables. These are represented by $Z_t$. Here the interpretation of $b_1$ is the effect on house price growth of a one unit change in the current account, whilst holding the control variables constant.

Models (3) and (4) address potential reverse causality. This is where changes in the dependent variable may cause changes in the independent variable, and as a result estimated coefficients are biased (Stock and Watson, 2012). These models use a one period lagged value of all independent variables. It is improbable that changes in the dependent variable in a period influence the independent variable in the previous period. Changes in the independent variables are made exogenous of changes in the dependent variable, eliminating bias.

As outlined in Section 2.3, the theoretical framework behind these models of price is the theory of supply and demand. My models analyse house price growth over a year long period and as such they are short run and assume the supply of housing is fixed. I only consider factors that influence the demand for housing.

### 3.3 Hypothesis Test

The $b_1$ coefficient gives the effect on house price growth of a unit change in the current account. I will use a hypothesis test to analyse the estimates of this coefficient, to assess whether the relationship between the two variables is statistically significant. Below is the test I will be conducting on the $b_1$ estimate for each model:

Null Hypothesis ($H_0$): $b_1 = 0$ \hspace{1cm} (5)

Alternative Hypothesis ($H_1$): $b_1 < 0$ \hspace{1cm} (6)

My choice to run a one-sided test originates from Bernanke’s (2005) hypothesis regarding the effects of the current account on house price growth, which says current account deficits are associated with
higher house price growth. Therefore the expectation is that negative changes in the current account result in positive house price growth, implying a negative coefficient. If I can reject $H_0$ at an acceptable confidence level, I can lend empirical support to this theory.

3.4 Variable Selection

The key variables in the models are those representing house price growth and the current account. $HPG_t$ gives year-on-year percentage changes in real house price growth, measured quarterly for period $t$. I use real rather than nominal growth because it removes inflation effects that distort the extent of growth.

$CA_t$ represents the year-on-year change in the current account balance which is expressed as a percentage of GDP, also measured quarterly for period $t$. I analyse the effect of changes in the current account balance rather than the absolute value of it because it reduces the problem of non-stationarity. This is where trends in the data cause biased coefficients and spurious regressions, giving a false representation of the relationship between two variables (Stock and Watson, 2012). This is also consistent with previous research.

$Z_t$, expressed as Equation (7), is comprised of five variables. I will now describe these, outlining what they represent and why they have been included. Details of measurement and data sources are included in Table A1 in the Appendix.

$$Z_t = \left[ GDP_t, \ INF_t, POP_t, INT_t, REC_t \right]$$

$GDP_t$ represents the year-on-year percentage change in real GDP per capita, measured quarterly. Changing income levels affect house prices through the demand side by changing people’s capacity and ability to borrow (Tsatsoronis and Zhu, 2004; and Igan and Loungani, 2012). Economic theory predicts a positive coefficient on this variable.

$INF_t$ represents inflation, the year-on-year percentage change in the price level, measured quarterly by the Consumer Price Index (CPI). Inclusion is based on the positive relationship between inflation and house prices identified by Tsatsoronis and Zhu (2004). The coefficient on this variable should be positive.

---

$^2$ See Bernanke (2010), Ferrero (2012)
POP\textsubscript{T} represents the year-on-year percentage change in the total US population, measured quarterly. Changes in population affect the demand for housing due to its function as shelter. For a fixed housing stock, a growing population should increase prices through an increase in demand. Both Tsatsoronis and Zhu (2004) and Igan and Loungani (2012) find demographic changes are an important driver of house prices. The coefficient on this variable should be positive.

INT\textsubscript{T} gives the year-on-year percentage change in the 30-year mortgage rate, a quarterly figure derived by averaging monthly data. Changes in the cost of borrowing affect house price growth through the demand side (Tsatsoronis and Zhu, 2004). A fall in the interest rate should encourage higher house price growth, suggesting the coefficient on this variable should be negative.

REC\textsubscript{T} is a dummy variable and indicates whether the economy was in recession in the respective period. This is defined as “a significant decline in economic activity spread across the economy lasting more than a few months” (NBER, 2010). This variable is included to account for the finding that house prices are pro-cyclical (Hirata et al., 2013). A recession should lower house price growth so the predicted sign of this coefficient is negative.

3.5 Sample Selection

My aim is to analyse the relationship between the two variables over a long time period to eliminate small sample bias. This will incorporate the most recent data available in order to ensure that my research augments existing literature and has policy relevant conclusions.

I have collected quarterly data on house price growth from Q4 1977 to Q4 2013 and it is over this period I will conduct my analysis. A larger sample would have been preferred. This could have been created by combining different datasets. However, I rejected this option due to differences in collection methods. Using monthly data would have enabled me to add more observations, but I have been unable to find data in this form. Each model will be run using 145 observations.

I am conducting a single country study focused on the US. Although findings of cross-country studies are correct on average, they may not apply equally to all countries in the study. Due to heterogeneity between countries it is difficult to use findings to inform policy. Furthermore, much of the cross-country data on property prices lacks comparability due to differences in collection methods (BIS, 2014a).
3.6 Data Description

Table 1 displays summary statistics for real house price growth and the current account balance.

**Table 1: Summary Statistics – Real House Price Growth and Current Account Balance (US, Q4 1977-2013)**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Real House Price Growth (%)</th>
<th>Current Account Balance (% of GDP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.49</td>
<td>-2.45</td>
</tr>
<tr>
<td>Median</td>
<td>1.33</td>
<td>-2.34</td>
</tr>
<tr>
<td>Maximum</td>
<td>13.74</td>
<td>0.66</td>
</tr>
<tr>
<td>Minimum</td>
<td>-18.66</td>
<td>-6.21</td>
</tr>
<tr>
<td>Range</td>
<td>32.4</td>
<td>6.87</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>6.44</td>
<td>1.67</td>
</tr>
</tbody>
</table>

Source(s): BIS (2014), Federal Reserve Bank of St. Louis (2014)

Note: Author’s own calculations

The data on the current account comes from the Federal Reserve Bank of St. Louis and has been seasonally adjusted. I use the current account balance as a percentage of GDP rather than the absolute value of the balance because it makes the data more comparable over time. On average the US has had a current account balance of -2.45% of GDP. The largest deficit recorded was in Q4 2005 where the figure was -6.20%. The standard deviation over the period was 1.67. Using the relationship between the current account and capital flows identified by Bernanke (2005), on average, the US has had positive capital inflows.

The data on house price growth is from the BIS database on property prices. It gives real, year-on-year, price growth for all existing dwellings, measured quarterly. When I refer to house price growth I am not referring solely to ‘houses’, the term includes similar goods such as flats which perform equivalent functions. This database is compiled using data provided by Central Banks, and the quarterly data has been derived by averaging monthly observations and deflated using the CPI (BIS, 2014a). Average house price growth over the period has been 1.49%. However the rate of price growth has varied widely with a standard deviation of 6.44. The highest house price growth was recorded in Q2 2005 at 13.74%; conversely, the biggest fall in house prices was observed in Q3 2008 at -18.66%. House price growth has been significantly more volatile than the current account over the period.
Figure 1 shows how the two variables have changed together over the sample period. From this a possible link between them emerges. Figure 1 is decomposed into two separate charts in Figures A1 and A2 in the Appendix.

Figure 1: Real House Price Growth and Current Account Balance (% of GDP) (US, Q4 1977 – Q4 2013)

Figure 1 shows the US has run a current account deficit over the period, apart from two periods from Q3 1980 to Q4 1981, and Q2 1991 to Q3 1991. House price growth is cyclical, with periods of price growth followed by price corrections. The length of these cycles varies. There was a strong period of growth between Q3 1983 and Q3 1987, followed by a sharp decline. There is then a longer period of price appreciation between Q2 1997 and Q2 2006. The fall in house price growth at the end of this cycle was severe with growth turning distinctly negative. The severity of this fall in growth is apparent by comparing it to previous periods of price correction where the rate of price growth never fell below -6.66%. In the house price appreciation phases, the current account balance is declining. When the current account moves further into deficit, house prices grow faster. For example, in the most recent appreciation cycle between Q2 1997 and Q2 2006 the current account balance declined from -1.35% to -5.81% of GDP. Similarly, falling house prices are associated with the current account balance moving upwards towards zero. This apparent correlation between the two variables furthers the case for empirical study.
3.7 Summary

In this Section I outlined the econometric models I will use to conduct my analysis of the impact of the current account on house price growth. These are constructed in a way that addresses omitted variable bias and reverse causality.

I also outlined the hypothesis test that will be used to investigate the underlying theoretical framework, which has testable implications. This is the theory that current account deficits are associated with higher house price growth.

I have explained my choice of variables, emphasising the decision to use real rather than nominal house price growth and changes in the current account rather than its absolute value. I have also described the theoretical rationale behind my choice of control variables. I chose to conduct a single country study over a long time period to ensure that results are policy relevant and that small sample bias is avoided.

Finally, I highlighted the persistent current account deficit run by the US, and the cyclicality of US house price growth. By running a formal econometric investigation, I will be in a better position to state whether the apparent relationship between the two variables is anything more than correlation.

4. Empirical Results

4.1 Introduction

Section 4 presents and analyses the results of my econometric investigation. Section 4.2 contains the estimated coefficients obtained after running each of my models. Here I describe my results and comment on aspects of the regressions such as model fit.

Section 4.3 discusses the result of the hypothesis test outlined in Section 3.3. This is done to assess the statistical significance of estimated coefficients on the current account variable. Section 4.4 considers confidence intervals for these estimates to further assess robustness.

Section 4.5 presents a sensitivity analysis of my results. This is done by running a robust regression to correct for distortions caused by outliers. I assess the extent of distortions in my original results by comparing estimated coefficients.
Section 4.6 further addresses reverse causality. To do this I use instrumental variable regression, utilising the Two Stage Least Squares approach (TSLS) and the more precise Generalised Methods of Moments (GMM) estimation. I compare results obtained using these approaches to those obtained using OLS regression. If they are similar, reverse causality is unlikely to be an issue.

Section 4.7 considers the limitations of my research and the conclusions I can draw from them. These limitations are used to identify future areas of investigation that would build on my contribution.

4.2 Results

Table 2 contains the coefficient estimates for each of my models, computed using robust standard errors to account for heteroscedasticity. Coefficient estimates computed with normal standard errors are contained in Table A2 in the Appendix.

**Table 2: OLS Regression Results**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>-3.15**</td>
<td>-1.59**</td>
<td>-3.04**</td>
<td>-1.59**</td>
</tr>
<tr>
<td></td>
<td>(0.64)</td>
<td>(0.61)</td>
<td>(0.61)</td>
<td>(0.68)</td>
</tr>
<tr>
<td>GDP</td>
<td></td>
<td>0.68**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.30)</td>
<td></td>
<td>(0.30)</td>
</tr>
<tr>
<td>INF</td>
<td>0.15</td>
<td></td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td></td>
<td>(0.17)</td>
<td></td>
</tr>
<tr>
<td>POP</td>
<td>-4.73*</td>
<td></td>
<td>-4.39*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.61)</td>
<td></td>
<td>(2.71)</td>
<td></td>
</tr>
<tr>
<td>INT</td>
<td>0.037</td>
<td></td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td></td>
<td>(0.04)</td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>-5.44**</td>
<td>-5.20**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.31)</td>
<td>(2.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>1.33</td>
<td>5.30*</td>
<td>1.31</td>
<td>4.83</td>
</tr>
<tr>
<td></td>
<td>(0.51)</td>
<td>(3.12)</td>
<td>(0.51)</td>
<td>(3.35)</td>
</tr>
<tr>
<td>N</td>
<td>145</td>
<td>145</td>
<td>145</td>
<td>145</td>
</tr>
<tr>
<td>$\bar{R}^2$</td>
<td>0.1297</td>
<td>0.3110</td>
<td>0.1200</td>
<td>0.2707</td>
</tr>
<tr>
<td>RMSE</td>
<td>6.0248</td>
<td>5.3605</td>
<td>6.058</td>
<td>5.5149</td>
</tr>
</tbody>
</table>

Note: ( ) are robust standard errors; * denotes significance at the 90% confidence level; ** denotes significance at the 95% confidence level.
To compare measures of model fit I will use the adjusted $R^2$ ($\overline{R^2}$). The $R^2$ measure is upwardly biased when multiple regressors are included (Stock and Watson, 2012). In Model (1) changes in the current account explain 12.97% of the variation in house price growth. The coefficient on the current account variable is negative and statistically significant at the 95% level. This implies negative changes in the current account balance raise the level of house price growth.

Model (2) explains 31.10% of the variation in house price growth over the period. The estimated coefficient on the current account variable falls by just below half; but remains negative and statistically significant at the 95% level. Again this suggests negative changes in the current account have a positive effect on house price growth. Both changes in real GDP per capita and whether there was a recession in the period have statistically significant effects on house price growth, the former positive, the latter negative. Changes in inflation are positively associated with house price growth and the effect of the interest rate is close to zero. However, both of these coefficients are not statistically significant. Positive population growth is associated with negative changes in house price growth, contradicting the theoretical prediction. A possible explanation for this is that it takes more than a year for demographic changes to affect house prices. An F-test has been used to test the hypothesis that all coefficients in Model (2) are equal to zero. This can be rejected at the 99% confidence level, suggesting the independent variables do have explanatory power.

In Model (3), using a lagged value of the current account variable explains 12% of the variance in house price growth. Again the estimated coefficient is negative and statistically significant at the 95% level, with a magnitude similar to that obtained for Model (1).

Using lagged values of all independent variables in Model (4) explains 27.07% of the variation in house price growth. The estimated coefficient on the current account variable is the same as that for Model (2) and is statistically significant at the 95% level. The estimated coefficients for the other regressors are also similar to those obtained for Model (2). Both GDP per capita and whether or not there was a recession exerts significant effects on house price growth. The effects of changes in inflation and the interest rate are not significant and the estimated coefficient on the population variable contradicts the theoretical prediction. The fact that the economic interpretation of the estimated coefficient on the current account variable is unchanged when lagged values are used suggests reverse causality is unlikely. Again an F-test has been used to test the hypothesis that all coefficients in Model (4) are equal to zero. This hypothesis can again be rejected at the 99% confidence level.
Model (2) is the preferred of the four models. It contains four statistically significant coefficients and given the variation in the $R^2$, Models (1) and (3) may suffer from omitted variable bias. Model (2) is preferred to Model (4) because it explains more of the variance in house price growth and minimises the Root Mean Squared Error (RMSE).

4.3 Testable Hypothesis

As specified in Section 3.3, I am running a hypothesis test to assess whether or not the estimated coefficients on the current account variable are statistically significant. I have found that all estimated current account coefficients are negative and statistically significant at the 95% level. Using critical values from the standard normal distribution, I am able to reject the null hypothesis that the value of these coefficients is zero. This adds support to research which argues that current accounts affect house prices, specifically that current account deficits are associated with higher house price growth (Bernanke, 2005).

4.4 Confidence Intervals

Table 3 contains 95% confidence intervals for the estimated $b_1$ coefficients in my models.

<table>
<thead>
<tr>
<th>Model</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>-4.42 ≤ $b_1$ ≤ -1.88</td>
</tr>
<tr>
<td>(2)</td>
<td>-2.80 ≤ $b_1$ ≤ -0.38</td>
</tr>
<tr>
<td>(3)</td>
<td>-4.24 ≤ $b_1$ ≤ -1.84</td>
</tr>
<tr>
<td>(4)</td>
<td>-2.95 ≤ $b_1$ ≤ -0.24</td>
</tr>
</tbody>
</table>

Note: Own calculations

A 95% confidence interval is the set of values that “has a 95% probability of containing the true value of $B_1$” (Stock and Watson, 2012, p. 193). None of the derived intervals contain the value zero, suggesting it is unlikely this is the true value of $b_1$. Therefore, it is unlikely that changes in the current account have no effect on house price growth.

4.5 Sensitivity Analysis

I have conducted a sensitivity analysis to assess the potentially distortionary impact of outliers in my sample. To do this I have run a robust regression for each of my models. Robust regression adjusts the
sample to exclude outliers or give them a reduced weight in the model, thereby reducing distortions to coefficient estimates (StataCorp, 2013a). It calculates Cook’s Distance to identify outliers before implementing two weighting functions, Huber and bi-weight, based on residuals (StataCorp, 2013b). Coefficient estimates derived from this methodology go beyond OLS regression in terms of their robustness in the presence of observations influenced by gross errors (Huber, 1964). To determine the extent of any distortion to my estimates I will compare the estimated coefficients for the current account variable with those obtained using OLS regression. Results are shown in Table 4.

Table 4: Sensitivity Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimated $b_1$, Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS Regression</td>
</tr>
<tr>
<td>(1)</td>
<td>-3.15**</td>
</tr>
<tr>
<td>(2)</td>
<td>-1.59**</td>
</tr>
<tr>
<td>(3)</td>
<td>-3.04**</td>
</tr>
<tr>
<td>(4)</td>
<td>-1.59**</td>
</tr>
</tbody>
</table>

Note: * denotes significance at the 90% confidence level; ** denotes significance at 95% confidence level

Table 4 shows there are differences in coefficient estimates. In Models (1), (3) and (4) the coefficient is revised down, in Model (2) it is revised up. Differences between estimates suggest there is some distortion from outliers in my original results. However, the average difference is 0.21, which, considered as a proportion of the coefficient, is small. In each robust regression the number of observations remained 145, meaning no observations were excluded. Even with revisions the estimated coefficients remain negative and statistically significant. Their economic interpretation remains unchanged.

4.6 Addressing Reverse Causality

To further assess potential reverse causality, I have utilised an instrumental variable approach. I use lagged values of both the change in the current account balance and the percentage change in population as exogenous instruments for the current account variable. This means the prediction for this variable is independent of the error term in the main regression and exogenous of changes in house price growth.
Model (5) is a basic model similar to Model (1). Model (6) includes the same control variables as Model (2). The only difference is the use of exogenous instruments to predict the current account variable. Models (7) and (8) are replicas of (5) and (6) respectively, but here the regression has been run using the GMM approach. This is a more precise method of instrumental variable regression that accounts for complications that effect the consistency and accuracy of estimators (Hansen, 1982). Results are displayed in Table 5.

Table 5: Instrumental Variable Regression Results

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Model (5)</th>
<th>Model (6)</th>
<th>Model (7)</th>
<th>Model (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>-3.88**</td>
<td>-2.15**</td>
<td>-4.13**</td>
<td>-1.74**</td>
</tr>
<tr>
<td></td>
<td>(0.87)</td>
<td>(0.97)</td>
<td>(0.80)</td>
<td>(0.81)</td>
</tr>
<tr>
<td>GDP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.56*</td>
<td>0.73**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.33)</td>
<td>(0.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
<td>(0.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4.94*</td>
<td>-4.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.68)</td>
<td>(2.60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-5.58**</td>
<td>-4.07*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.66)</td>
<td>(2.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>1.29**</td>
<td>5.67*</td>
<td>0.877*</td>
<td>5.24*</td>
</tr>
<tr>
<td></td>
<td>(0.50)</td>
<td>(2.87)</td>
<td>(0.47)</td>
<td>(3.11)</td>
</tr>
<tr>
<td>N</td>
<td>145</td>
<td>145</td>
<td>145</td>
<td>145</td>
</tr>
<tr>
<td>Centred R²</td>
<td>0.1284</td>
<td>0.3366</td>
<td>0.1186</td>
<td>0.3356</td>
</tr>
<tr>
<td>Sargan Statistic</td>
<td>3.811</td>
<td>3.063</td>
<td>4.162</td>
<td>2.980</td>
</tr>
<tr>
<td>Chi Squared (1) p-value</td>
<td>0.0509</td>
<td>0.0801</td>
<td>0.0413</td>
<td>0.0843</td>
</tr>
</tbody>
</table>

Note: ( ) are normal standard errors; * denotes significance at the 90% confidence level; ** denotes significance at the 95% confidence level.

To assess whether my instrumental variable approach is valid I analyse the Sargan Statistic and the associated p-value. For Models (5), (6) and (8), the p-value is greater than 0.05, meaning I can accept
the null hypothesis that my instruments are valid at the 95% confidence level. The p-value for Model (7) suggests that using the more precise GMM approach my choice of instruments may not be valid.

To compare the coefficient estimates obtained using this approach to those obtained through OLS regression, I have conducted a Hausman Test. This compares estimates under the null hypothesis that there are no significant differences between them, acceptance of which implies that both are good estimates of the true parameters (Hausman, 1978). When each TSLS model is compared to its OLS counterpart I am able to accept this null hypothesis at the 1% significance level.

I have been able to generate results using the instrumental variable approach similar to those obtained using OLS regression. Controlling for reverse causality, estimated coefficients on the current account variable are negative and statistically significant at the 95% level, with the same economic interpretation. This suggests reverse causality is not a problem in my original results.

4.7 Limitations

A weakness of my results and the conclusions I draw from them is that my models fail to explain all of the variance in house price growth. The highest proportion of variance accounted for by my core models is 31.10%. It is possible my models suffer from omitted variable bias and that the inclusion of additional variables may affect the magnitude, direction and interpretation of the coefficients on the current account variable.

Such omitted variables could include the supply of housing. I chose not to include this variable because I considered changes in price over a short run period. However, in the long run changes in supply are an important determinant of price and the incorporation of these changes into a similar analysis would add value to the literature.

Another variable not accounted for is structural breaks, for example a government policy that affects house prices for a portion of the sample. Future research could analyse government housing policies and classify them as having a positive or negative effect on prices. This information could be included as a dummy variable in an econometric framework similar to that used in this dissertation.

A further limitation is that I have not analysed the channels through which changes in the current account affect house prices. I have established a relationship between the two variables, however, further investigation is needed to identify why this is the case. Of concern is that the channel through
which the current account influences house prices, identified by previous research, is through its effect on interest rates. However, in my results the estimated coefficients on the interest rate variable is close to zero and not statistically significant. This implies changes in this variable have little impact on house price growth. This may be due to my decision to use the 30-year mortgage rate to represent the interest rate in my model. An alternative measure may have yielded a different result.

Finally, because I have focused on a single country, my results lack external validity. My conclusions are relevant to the US, but may not be to other countries, especially those with significantly different housing markets and economies generally. It would be interesting see whether my results could be replicated by future research, using data from other countries.

4.8 Summary

I have found that negative changes in the current account were associated with higher real house price growth in the US between Q4 1977 and Q4 2013. This effect is significant even when controlling for other factors that influence house prices. I have shown that the estimated coefficient on the current account variable is, in all of my models, statistically significant and unlikely to equal zero.

Distortions to my estimated coefficients due to outliers are minimal. Estimates are slightly different under robust regression, though, this difference is small and the results remain statistically significant and of similar direction and magnitude.

Throughout my analysis I have addressed reverse causality. I have made use of lagged variables and instrumental variable regression to make changes in the current account exogenous of changes in house prices. The results generated using these methods suggest reverse causality is not an issue in my main results.

Whilst my findings add value to the existing literature, the impact of my conclusions is limited. My models fail to account for all of the variance in house price growth, suggesting they may suffer from omitted variable bias. I have also been unable to account for structural breaks in my sample and changes in housing supply.
5. Conclusion

It is important to understand the housing market, both because of the role housing plays in society and the externalities that can arise from it. These are particularly relevant for the financial sector.

There is general agreement regarding the core determinants of house prices. However, there is disagreement regarding the relationship between the current account and house prices. Initial research suggested that current accounts drive house prices, however, later research claimed the relationship is the other way around. In a globalised world, understanding the impact of cross-border flows will only become more important.

The importance of the housing market and the lack of clarity in the existing literature is why I felt it important to empirically investigate the impact of the current account on house prices. I sought to do this in a robust way that would fill a gap in the existing literature. Crucial to achieving this was to accept and address the potential for reverse causality and study the relationship between the two variables over an extended period.

To do this, I conducted an econometric analysis, focusing on the US over a 37-year period. I used four model specifications aimed at addressing the weaknesses in the existing literature and conducting an investigation that would produce robust results. I focused on minimising omitted variable bias and controlling for potential for reverse causality.

The results of my analysis suggest negative changes in the current account are associated with higher rates of house price growth. A shift towards a current account deficit leads, ceteris paribus, to higher house price growth. This finding holds even when controlling for changes in variables that have been shown to affect house price growth.

This is an important finding because it informs one as to the consequences of running a current account deficit. There are also implications for countries looking to adjust their current account position. My findings inform policymakers as to the potential spill over effects on to the housing market of current account adjustment. As changes in the housing market produce externalities linked to the wider financial market and the path of inequality, understanding the effects of adjustment on the housing market is important for ensuring awareness of potential unintended consequences in these areas.
There are a number of limitations to my analysis that detract from my conclusions. My models do not account for all of the variance in house price growth over the period, which suggests there are factors, which I have not included in my model that should be. Furthermore, because I have conducted a single country study, my findings lack external validity and may not be applicable to other countries, especially those significantly different to the US.

To conclude, I have found that negative changes in the current account balance were associated with higher levels of house price growth in the US between Q4 1977 and Q4 2013. This finding holds even when controlling for changes in real GDP per capita, inflation, population, interest rates and the state of the economy. My findings are empirically robust and my study has been conducted over a significantly longer time period than the majority of existing literature. Unlike much of the previous research I have directly addressed and accounted for potential reverse causality and my results suggest this is unlikely.
Bibliography


### Table A1: Regression Variable Descriptions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Frequency</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_P^G$</td>
<td>Percentage change in real house prices for all dwellings (US) (YoY)</td>
<td>Quarterly</td>
<td>Bank for International Settlements (2014b)</td>
</tr>
<tr>
<td>$C_A$</td>
<td>Change in Current Account Balance as a percentage of GDP (US) (YoY)</td>
<td>Quarterly</td>
<td>Federal Reserve Bank of St Louis (2014)</td>
</tr>
<tr>
<td>$G_P^D$</td>
<td>Percentage change in Real GDP per capita (US) (YoY)</td>
<td>Quarterly</td>
<td>Federal Reserve Bank of St Louis (2015a)</td>
</tr>
<tr>
<td>$I_N^F$</td>
<td>Percentage change in the Consumer Price Index (US) (YoY)</td>
<td>Quarterly</td>
<td>Federal Reserve Bank of St Louis (2015b)</td>
</tr>
<tr>
<td>$P_O^P$</td>
<td>Percentage change in total population (US) (YoY)</td>
<td>Quarterly</td>
<td>Federal Reserve Bank of St Louis (2015c)</td>
</tr>
<tr>
<td>$I_N^T$</td>
<td>Percentage Change in 30-Year Mortgage Rate (US) (YoY)</td>
<td>Monthly (Averaged over the Quarter)</td>
<td>Freddie Mac (2015)</td>
</tr>
<tr>
<td>$R_E^C$</td>
<td>Dummy variable. Rec = 1 if the economy was in a recession during the time period (US)</td>
<td>Quarterly</td>
<td>National Bureau of Economic Research (2010)</td>
</tr>
</tbody>
</table>

Note: YoY = Year-on-Year
Figure A1: Current Account Balance (% of GDP) (US, Q4 1977 – Q4 2013)

Source(s): Federal Reserve Bank of St Louis (2014)

Figure A2: Real House Price Growth (%) (US, Q4 1977 – Q4 2013)

Source(s): BIS (2014b)
<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Model (1)</th>
<th>Model (2)</th>
<th>Model (3)</th>
<th>Model (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>CA</td>
<td>-3.15** (0.67)</td>
<td>-1.59** (0.69)</td>
<td>-3.04** (0.67)</td>
<td>-1.59** (0.71)</td>
</tr>
<tr>
<td>GDP</td>
<td></td>
<td>0.68** (0.31)</td>
<td></td>
<td>0.63** (0.32)</td>
</tr>
<tr>
<td>INF</td>
<td>0.15 (0.21)</td>
<td>0.17 (0.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POP</td>
<td>-4.73* (2.73)</td>
<td></td>
<td>-4.39 (2.83)</td>
<td></td>
</tr>
<tr>
<td>INT</td>
<td>0.04 (0.05)</td>
<td></td>
<td>-0.02 (0.05)</td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>-5.44** (1.69)</td>
<td></td>
<td>-5.20** (1.74)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>1.33 (0.50)</td>
<td>5.30** (2.90)</td>
<td>1.31 (0.51)</td>
<td>4.83 (3.06)</td>
</tr>
<tr>
<td>N</td>
<td>145</td>
<td>145</td>
<td>145</td>
<td>145</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.13</td>
<td>0.31</td>
<td>0.12</td>
<td>0.27</td>
</tr>
<tr>
<td>RMSE</td>
<td>6.0248</td>
<td>5.361</td>
<td>6.058</td>
<td>5.5149</td>
</tr>
</tbody>
</table>

Note: ( ) are normal standard errors; * denotes significance; ** denotes significance at the 95% confidence level.
The Effect of the UK’s Smoke Free Law Has Had on Smoking Prevalence, Tobacco Related Hospital Admissions and NHS Expenditure – A Discussion Seven Years following the Public Smoking Ban

Kate Wales

The consumption of tobacco is a widely researched and discussed area, particularly in terms of health effects. This dissertation aims to add to current and past literature by discussing the impact the smoke free legislation has had on smoking prevalence, tobacco related hospital admissions and consequently NHS expenditure on tobacco related diseases. Particular analysis will be given in terms of age, gender and occupational classification. The methodology and data used is discussed, giving limitations and reference to the quality. The results present the data between years 1974 and 2013, giving interpretation to the trends found. In conclusion, this dissertation finds that smoking prevalence has been largely uninfluenced by the introduction of the smoking ban and hospital admissions due to tobacco related diseases are yet to decline.
1. Introduction

There is extensive research and literature on the health effects of tobacco and consumption of cigarettes. There is additional research on how different tobacco control measures have influenced health and consumption, however, there are limited studies that focus on the effects the smoking ban in the UK had on tobacco consumption and National Health Service (NHS) expenditure. Therefore, the first objective of this paper is to analyse how trends in smoking prevalence have changed pre- and post-smoking ban. Secondly, the number of tobacco related hospital admissions will be assessed in light of the smoking ban, and the final objective is to study whether the smoking ban has impacted NHS expenditure on such diseases. The economic burden of tobacco has been estimated to cost society £13.74 billion per year, this includes: loss of productivity, healthcare costs, cessation services and help from councils in cleaning up (Department of Health, 2011). A combination of tobacco control measures, such as imposing higher taxes and restricting public smoking, is needed in order to reduce tobacco consumption and the negative health effects (Jones et al., 2015). Raising tobacco taxes was used primarily to deter consumption, whereas the introduction of the smoking ban in 2007 aimed to reduce the harmful effects of Second Hand Smoke (SHS). It has been shown that raising taxes has been one of the most effective means of deterring consumption amongst the younger and less well off people in society (Gilmore et al., 2013), and the smoke-free law has reduced the number of incidents of hospital admissions due to SHS (Department of Health, 2011). Various cohorts within society are affected differently. Jones et al. (2015) acknowledges there will be differential impacts from the smoking ban, dependant on age, gender and pre-ban levels of tobacco consumption.

This paper aims to isolate the effect the smoking ban has had on smoking prevalence, hospital admissions and in turn NHS expenditure on tobacco related diseases. To achieve this, I will firstly review current and past literatures to assess changing consumption patterns of tobacco, associated cost to the NHS and appraise the success of previous tobacco control measures. Section 3 will then go on to discuss the methodology used in this study, evaluating the data and any limitations associated with it. Following this, the results are presented in Section 4, with particular focus on age, gender, occupational classification and hospital admissions. The results are discussed and interpreted accordingly. Section 5 concludes the findings of this study including any limitations found. It states that smoking prevalence has been largely unaffected by the introduction of the smoking ban, however, the incidence levels from SHS exposure, especially amongst children, has fallen. Hospital admissions due to tobacco related diseases are yet to decline. Finally this paper suggests that further research needs to be undertaken to accurately assess the monetary burden tobacco consumption has on the NHS.
2. Literature Review

The aim of this Section is to provide an overview of the academic literature that assesses the impact different tobacco control measures have had on both cigarette consumption and NHS expenditure, with emphasis on the smoking ban. A focus will be placed on different cohorts; gender, age and occupational group. This Section will provide evidence from previous studies to examine these effects.

2.1 Overview of tobacco consumption within England

Tobacco consumption is a major preventable cause of death in countries all over the world (World Health Organisation (WHO), 2004). Specifically, within the UK, the highest impact on the NHS is through behavioural factors; poor diet, alcohol consumption and smoking (Scarborough et al., 2011). This insinuates that, through national policies aiming to alter individuals’ behaviours, such as a smoking ban, a substantial saving to the NHS can be made. WHO (1998) estimated worldwide that tobacco would be responsible for 10 million deaths per year between the decades 2020 – 2030. Consequently policies to reduce tobacco consumption have been a major focus for all countries.

The UK follows similar tobacco consumption patterns to other developed nations. Specifically, within England, deaths from tobacco account for more than the next six most preventable causes of death combined, which includes drug use, road accidents, other accidents and falls, preventable diabetes, suicide and alcohol abuse (Department of Health, 2011, p.5). In 2009 smoking accounted for 81,400 deaths and cost the NHS approximately £50 million per week (Department of Health, 2011, p.5). This shows that tobacco consumption is a huge problem within the UK, both to the consumer, secondary consumers through SHS (secondhand smoke) and to the nation financially. From the mid 20th Century onwards, varying degrees of tobacco consumption control programs have been implemented, such as: restricting advertising, regulation measures, cessation services, fiscal policy of taxation and the national smoking ban. The US Surgeon General stated that:

“the mission of comprehensive tobacco control programs is to reduce disease, disability, and death related to tobacco use, a comprehensive approach has been established as the guiding principle for eliminating the health and economic burden of tobacco use.” (US Department of Health and Human Services, 2000)

The most influential and widely debated measures include fiscal policy changes and the national ban, both discussed in detail as follows:
2.2 Smoking ban on Tobacco Consumption

The ban on smoking in public places, as a measure to reduce consumption of tobacco, has been favoured not only from the health care specialists and policy makers but also by the public in large. Chollat-Traquet (1996) acknowledged that among the public there is an agreed consensus that they have the right to breathe air that is not contaminated with the harmful effects of tobacco smoke. Since July 2007, smoking has been prohibited throughout the UK in all enclosed and substantially enclosed work and public places. A ‘substantially enclosed’ area means premises or structures with a ceiling or roof (including retractable structures such as awnings) and where there are permanent openings, other than windows or doors, which in total are less than half of the area of walls (Health Act 2006). This policy was implemented in order to create a smoke free workplace and smoke free public areas, whilst encouraging cessation among current smokers.

Tobacco smoke has proven to be carcinogenic to humans (International Agency for Research on Cancer, 2004) and contains 43 known cancer-causing agents (WHO, 1998), thus, causing the harmful effects of SHS. Scientific evidence has shown that exposure to SHS causes death, disease and disability (WHO, 2005). The primary aim of the smoking ban was to protect non-smokers from the harmful effects of SHS, with a secondary intention of encouraging current smokers to quit or to reduce the consumption of cigarettes (Jones et al., 2015). The smoking ban has proven to be successful in reducing the number of hospital admissions for heart attacks due to a smoke free environment (Bauld, 2011). There has also been a 70% reduction between 1996 and 2007 in SHS exposure among children, which may partially be attributable to media campaigns leading up to the smoking ban (Bauld, 2011). The Health and Social Care Information Centre (HSCIC) have also published data showing an 18% decrease in smoking related hospital admissions among adults aged 35 and over between the years 2004/05 and 2012/13. The percentage of deaths caused by smoking, however, has been unchanged since 2005 (Lifestyle Statistics, Health and Social Care Information Centre, 2014).

Smoking and consumption patterns differ between cohorts, causing policies to have a higher impact within specific groups. Tobacco consumption is highest among lower socio-economic groups, typically in semi-skilled manual labour jobs, it is also most popular among 20-24 year olds; there is, however, a converging level of incidence of smoking between genders (Royal College of Physicians, 2000). Previous research of the effect the smoking ban has had on different cohorts is summarised below.
2.2.1 Impact of Smoking Ban Within Cohorts

Jones et al. (2015) evaluates the impact of the smoking ban in the UK on active smokers and discovers that there were limited short-run effects on both total level of smoking and smoking prevalence. The ‘short-run effect’ means that the smoking ban has a time lag effect. As it has only been in place for a limited number of years, the benefits from cessation may not be fully realised, thus prohibiting a long run measure.

a) Age and Gender

Jones et al. (2015) reveal that gender and age play a role in differing consumption patterns. Their results show that differences in the patterns of consumption are more evident in the cohort 18-34. Among males, those who do smoke do so more intensively but with fewer males smoking overall, whereas females within this cohort consume less cigarettes throughout their lifetime but are less likely to break the habit (Jones et al., 2015). When the smoking ban came into place, in 2007 there was a significant increase in the number of people who reported trying to quit, amounting to 300,000 (Glasper, 2011). Further studies have shown, however, that consumption of tobacco has not been affected by the smoking ban, indicating that those particular individuals failed to quit.

A study carried out by Jones et al. (2015) examined the success of the smoking ban in relation to the differential timing the legislation was legalised within England (1st July 2007) and Scotland (26th March 2006). It was quasi-experimental, exploiting variation over time and between different cohort groups, using data from the British Household Panel Survey. The main findings from the study were that the introduction of the smoking ban in both countries had limited effects on smoking prevalence and total level of smoking in the short run. Cohorts were affected differently, though. The results showed that the ban may have caused the number of cigarettes smoked among older men to decrease by 1.4 cigarettes a days, but among male moderate smokers consumption increased by 1.6 cigarettes a day; whilst the corresponding coefficients for females were not statistically significant (Jones et al., 2015).

Other figures have shown that across the English population, 26% of people aged 16-24 smoked in 2009; over recent years this rate has begun to decline, though (Department of Health, 2011). This supports the idea of the short run effect in the Jones et al. (2015) paper, as younger cohorts are deterred from smoking thus decreasing smoking prevalence levels in future years. Allender et al. (2009) agree that the smoking ban is subject to time lags, thus, having a limited short run effect on prevalence figures and hospital admissions, especially among younger cohorts.
b) Occupational Classification

Consumption patterns and cessation rates have been shown to vary dramatically between occupational groups. It has been found that there is a strong relationship between smoking and occupation, with prevalence rates being twice as high among those in routine and manual occupations to those in managerial and professional occupations (Department of Health, 2011). A study on the patterns and predictors of tobacco consumption among women found that in Britain cessation rates among the poorest groups has changed little across two decades whilst in the better off socio-economic cohorts, the cessation rate has more than doubled within the same time frame (Graham and Der, 1999). This is supported in a more recent study that claims “smoking-related death rates are two to three times higher in low-income groups than in wealthier social groups” and tobacco control programs are thus targeting this group due to the high tobacco consumption (Department of Health, 2011, p.17).

c) Pre Ban Average Daily Consumption of Cigarettes

Jones et al. (2015, p.190) states that “trends of smoking differ substantially according to the pre-ban average daily consumption”, indicating that the ban may only have a significant effect on heavy smokers consuming more than twenty cigarettes a day, and insignificant effect on lighter smokers. De Chaisemartin et al. (2011) and Anger et al. (2011) agree that the smoking ban hasn’t reduced smoking in the whole population; they do find, however, that consumption for social smokers, who frequently visit bars and restaurants, has been reduced. Callinan et al. (2010) provide further support that a smoking ban has little or no effect on active smokers, but they find that it does reduce the effects of passive smoking. The outcome of this study showed no consistent evidence of a decrease in smoking prevalence that was attributable to the ban, but did provide support for the argument that there was a reduction in hospital admissions for cardiac events as well as an improvement in general health as a result of reduced SHS (Callinan et al., 2010). Furthermore, Irvine et al. (2011), predict that, smokers may now consume their cigarettes within a shorter period of time, making their consumption more intense, than if they were able to smoke in any environment prior to the ban.

2.3 Fiscal Measures to Reduce Tobacco Consumption

Alongside the national smoking ban, the government has used fiscal measures of taxation on tobacco products to reduce the level of consumption. WHO (2004, p.37) stated that “for tobacco control to succeed, a comprehensive mix of policies and strategies is needed, if resources are limited efforts should focus first on raising tobacco prices through increased taxes”. Within the EU, the UK has some of the highest priced tobacco products available, as the government tries to deter purchases
Department of Health, 2011); this has been used in conjunction with other policies alongside the smoking ban to reduce consumption. WHO (2004) reported that for every 10% increase in cigarettes taxes, there is approximately a 4% reduction in consumption, which clearly demonstrates the inverse relationship between tobacco consumption and tobacco taxes.

There is an abundance of literature on the price elasticity of demand for tobacco, some of which is summarised here. As previously discussed the highest prevalence rates are among the younger and lower occupational groups, with WHO (2004) realising that minorities, young persons and low-incomes are the most responsive to price increases. This is further supported in a paper by Gilmore et al. (2013, p.1317), who state that “raising tobacco taxes and prices is one of the most effective means of reducing tobacco use, particularly in the young and the less well off – who are known to be the most price sensitive”. A study carried out in 2006 showed that under the impact of rising taxes, consumption of both duty-paid and smuggled tobacco has fallen in volume terms per smoker since 1990 (Duffy, 2006). Gilmore et al. (2013) look beyond the current changing consumption of tobacco in relation to price and determined an overshifting pricing strategy, where prices are increased on top of tax increases, causes consumers to move from a premium brand to a lower brand and may undermine the public health impacts of the tax increases. It is important to note that tobacco tax accounts for only 3.6% of government receipts (Royal College of Physicians, 2000), inferring that the government is primarily focused on the health implications of deterring tobacco consumption, as opposed to the monetary rewards. A study in the city of Jaipur, India supports these findings. The government of Rajasthan increased tax from 20-40% on all tobacco products from 1st April 2011, with the results summarising that on average a 10% increase in price of cigarettes led to a 8% reduction in consumption within that state (Singh et al., 2012). However, they also found that 64% of cigarette users consumed the same amount of tobacco (Singh et al., 2012), indicating that once consumers were already addicted price became less of a deterrent. This is further supported by a study in the US, as they conclude there is insufficient evidence to justify that raising cigarette taxes will significantly reduce cigarette consumption among adults and believe that at best a tax increase of 100% will decrease smoking by as little as 5% among smoking adults (Callison et al., 2014). Notably, the pool of adult smokers within this year are more likely to have a strong preference towards continuing to smoke, as a they have already been subjected to significant increases in tobacco taxation and continue to partake, but the study does, however, conclude that raising taxes on cigarettes gets a higher response rate from younger cohorts and deters consumption (Callison et al., 2014).
2.4 Smoking: Costs to the NHS

From what has been previously mentioned, it is irrefutable that the national smoking ban in 2007 was brought in to lessen the damaging effects of SHS and increase the cessation rate amongst smokers, thus improving the health of the nation. Due to smoking imposing such a large burden on population health and NHS resources, it is important to quantify this burden to help prioritise the limited NHS resources (Allender et al., 2009). It is apparent that there has been a fall in consumption of tobacco since the 1990’s due to a combination of tobacco control measures. This is reflected in NHS expenditure as the cost of smoking in 2006 was estimated to be 13% lower than if smoking had remained at 1996 levels (Callum et al., 2011). In 2006/07 alone, treating smoking-related illnesses cost the NHS approximately £2.7 billion, amounting to over £50 million every week (Department of Health, 2011). A comprehensive study carried out by Callum et al. (2011) broke down the costs to the NHS of tobacco attributable diseases in terms of consultations, overnight stays, prescriptions and costs of treating the disease, where the diseases caused by smoking are thought to be those with a causal link, i.e. not diseases which are exacerbated by smoking. The results show that, using 2006 unit costs, in 1996 the burden to the NHS amounted to £3.09 billion compared to £2.70 billion in 2006 which can be accounted for by the decrease in prevalence rates (Callum et al., 2011). This study is limited purely to diseases caused by individuals smoking and doesn’t take into account the effects of passive smoking, maternal smoking during pregnancy nor the cost to society of informal care and loss in productivity. Parkin (2011) estimated that 60,837 (19.4% of all new cancer cases) are attributable to tobacco; that is 36,537 (23%) of cancers in men and 24,300 (15.6%) of cancers in women, and 86% of total lung cancer diagnosis is due to exposure to tobacco smoke. It is reported that the smoking ban has contributed to fewer children being exposed to SHS at home, thus reducing preventable expenditure of the NHS (Department of Health, 2011).

2.5 Summary

It is evident that the smoking ban has had a minimal effect on smoking prevalence rates amongst the UK population. Smoking prevalence has decreased over the last decade; nevertheless the contribution of the smoking ban is minimal in comparison to fiscal measures. Despite, smoking fatalities being amongst the most preventable causes of death, there has been no notable decrease amongst smoking related diseases, particularly among those aged below 35. However, the level of incidence from SHS has decreased, specifically in the hospitality industry and among children. Prevalence rates are highest among the younger cohorts, i.e. those between 18-34 years of age, and in lower occupational groups, thus making them the target of new policies. It is to be noted that there is a significant time lag when
analysing hospital admissions and disease burden, as a sufficient number of years has not yet passed to calculate the long run effect of changing smoking habits.

3. Methodology

The aim of this Section is to provide a clear understanding of the methodology used in this paper, discussing the data and any limitations that arise from it. This paper is a comparative case study, looking at tobacco consumption patterns before and after the smoking ban came into place. It further explores the effect the ban has had on tobacco related hospital admissions and how it has, in turn, impacted NHS expenditure. The years 1974 to 2013 have been studied where possible, giving an extensive period for a time-series analyses to be conducted.

Time-series is used to provide a platform to extrapolate trends alongside a log-linear model to calculate trend growth both pre- and post-smoking ban. Furthermore, paired t-tests have been used to determine whether the null hypothesis, i.e. that the mean level of smoking prevalence is equivalent amongst both male and females, can be rejected or accepted. To support these quantitative findings, literature from reports, studies and other publications will be used in conjunction to provide both context and further evidence to any conclusions drawn.

The data for this study has been taken from two organisations, The Office of National Statistics (ONS) and The HSCIC. ONS published data from the Opinions and Lifestyle Survey (OPN), General Lifestyle Survey (GLF or sometimes referred to as the GLS) and the General Household Survey (GHS, formally known as the GLS). These surveys have provided continuous data on the economy, population and society from private households since 1971. The ONS is an independent producer of statistics, thus reducing the possibility of bias. Due to the surveys being conducted for over 40 years, survey questions have been adapted in terms of both content and design to make it more relevant to the year in question, however, not affecting the results received. The GLF and GHS have been continuously used for legislative purposes, to give public bodies a more accurate way of analysing new legislations, such as the smoking ban (Office for National Statistics, 2013b). HSCIC publishes data from both the Health Survey for England (HSE) and the Hospital Episode Statistics (HES), both of which are used in this study. The HSCIC is a non-departmental public body sponsored by the Department of Health that provides information and data about health and social care (Health and Social Care Information Centre, 2015). The HSCIC has been providing information to the NHS since 1991 allowing it to run as efficiently and effectively as possible (Health and Social Care Information Centre, 2015).
3.1 Measurement of the Proportion of Smokers within Different Cohorts

The OPN, GLF and GHS provided data for the different cohorts studied in this paper, namely age, gender and occupational classification. Both the OPN and GLF use a randomised method to select participating households, giving equal chance of each household being selected across Britain. From those selected there is only a 60% participation rate; this is, however, still representative of the nation after the sample has been weighted (Office for National Statistics, 2013c). Estimates from the year 2000 onwards are weighted, meaning the sample results are scaled to population size and demographics to provide an accurate representation of the whole population. There is a minor different in the survey approach of the OPN and GLF. The GLF interviews all members of the randomly selected households aged 16 and over, whereas the OPN just interviews one member aged 16 or over. Nevertheless, it has been shown there is no significant difference in results stemming from the slight variation in method.

3.2 Measurement of Current Smokers Perceptions of the Smoking Ban

Data obtained from the HSE has been used to demonstrate individuals’ perceived success of the smoking ban and the most favoured reasons for smokers wanting to quit, thus helping to isolate the impact of the smoking ban over other tobacco control measures. The HSE uses multi-staged stratified sampling to select households to conduct face-to-face interviews, self-completion questionnaires, clinical measurements, diaries, physical measurements and CAPI to obtain the information needed. In order to gain information on the ‘perceived success’ of the smoking ban, smokers were interviewed six months before and after the smoking ban came into place. This allowed disparities to be calculated between intentions and actions pre- and post-smoking ban.

3.3 Measurement of NHS Hospital Admissions and Primary Diagnosis

Data on the number of tobacco related hospital admissions was taken from the HES. Numerical estimates for the number of smoking-attributable admissions was calculated ‘using risk ratios for diseases associated with smoking-attributable fatalities…with additional risk ratios for non-fatal diseases attributable to smoking’ (Lifestyle Statistics, Health and Social Care Information Centre, 2013, p.83). It is important to note that the data included private patients in NHS hospitals, however, it didn’t account for private patients in private hospitals and only those that are resident in England.

3.4 Limitations

Throughout the surveys used, there is potential that some of the figures and statistics obtained have been underestimated. Firstly, on average, consumption is rounded down to the nearest five or ten
cigarettes consumed. Secondly, underestimation in prevalence arises between those aged 16 and 17 as individuals are influenced by the presence of other family members who may disapprove of cigarette consumption within their household. As a result the ‘smoking section’ of the interview is given as a self-complete questionnaire to complete in private. Finally private patients in private hospitals are excluded from the number of hospital admissions. In this study to be regarding as a heavy smoker, twenty or more cigarettes have to be consumed per day.

A significant limitation of this methodology is isolating the effect the smoking ban has compared to other tobacco control measures. Many factors contribute to changes in smoking prevalence such as finance, health, national policies and personal circumstances. Furthermore, determining whether a tobacco related disease is the primary cause from smoking or is merely exacerbated by smoking is difficult to differentiate.

4. Results and Interpretation

In this Section, a summary of time series data from years 1974-2013 on smoking prevalence between gender, age and occupation are discussed and interpreted, followed by a section on perceived reasons for wanting to quit smoking. This addresses the first objective stated in Section 1. Following this, NHS hospital admissions related to smoking will be presented and interpreted to address the second objective. Subsequently, to address the third objective, analysis will be undertaken to see if there is any correlation between the first two objectives and NHS expenditure.

Section 4.1: Statistics on the Proportion of Smokers by Gender and Age

Figure 4.1 shows the percentage of men who smoke cigarettes in Great Britain between 1974 and 2013. It is evident that there has been a substantial decline in smoking prevalence among men in all age groups since 1974. It becomes apparent that the group aged 60 years and over has had a consistently lower proportion of men smoking compared to any of the other age groups, whereas the cohort of 25-34 year-olds has had the highest smoking prevalence on average. The age group seeing the largest decline in smoking prevalence between 1974 and 2013 was the one classified as ‘aged 60 and over’, with a 72.73% decline, followed by ages 50-59, 35-49, 25-34 and, lastly, 16-24 with a 44.68% decrease. From when the smoking ban came into place throughout the UK in 2007, Table 4.1 shows percentage declines and average annual percentage declines for both pre- and post-smoking ban within cohorts. There has been only a minor decrease in smoking prevalence amongst men in cohorts post-ban, the largest difference being the 50-59 year olds with an average annual decrease of 1.52%. However, in accounting for all persons aged 16 and over there has been an average annual 0% change.
in proportion of men who smoke cigarettes between 2007 and 2013. Smoking prevalence is shown to be declining at a decreasing rate as the pre-ban growth rate was -3.5% compared to a post-ban rate of -0.2%. This has also been discovered in other studies, where it has been found that rates of smoking have declined over the past few decades, but that there has been a relatively small decrease in smoking prevalence since 2007 (Department of Health, 2011).

**Figure 4.1: Proportion of Men Who Smoke Cigarettes in Great Britain, 1974-2013**

Figure 4.2 shows the percentage of women who smoke cigarettes in Great Britain. Similarly to men, it is evident that there has been a substantial decline in smoking prevalence amongst women between 1974 and 2013, however, the peak prevalence rate in 1974 never exceeded 49%, unlike the male equivalent of 55%, both being in the 35-49 age group. Amongst women, the 60 and overs cohort shows that a significantly smaller proportion of females in that age group smoke cigarettes, never exceeding 26%. It is noteworthy that this same female cohort, 60 and overs, has the largest percentage decrease in smokers between the years 1974 and 2013, calculated to be 61.54%. Age group 50-59 closely follows this, and then 35-49, 25-34 respectively and lastly 16-24 with a 51.22% decrease. From 1986
onwards, the higher aged cohorts begin to show a sharper decline in smoking prevalence than those aged between 16 and 34. Smoking norms prior to the 1980’s, show that over 50% of men smoked and over 40% of women smoked specifically in the year 1972 (Royal College of Physicians, 2000), this coincides with a time when adverse health effects weren’t fully realised.

Table 4.2: Percentage Decline in Smoking Prevalence by Gender, Pre- and Post-Smoking Ban, Great Britain, 1974-2013

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Decline 1974 - 2006 (%)</th>
<th>Average Annual % decrease 1974 - 2006 (%)</th>
<th>Decline 2007 - 2013 (%)</th>
<th>Average Annual % decrease 2007 - 2013 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 - 24</td>
<td>44.68</td>
<td>1.40</td>
<td>3.70</td>
<td>0.62</td>
</tr>
<tr>
<td>25 - 34</td>
<td>40.00</td>
<td>1.25</td>
<td>-3.45</td>
<td>-0.58</td>
</tr>
<tr>
<td>35 - 49</td>
<td>52.73</td>
<td>1.65</td>
<td>4.00</td>
<td>0.67</td>
</tr>
<tr>
<td>50 - 59</td>
<td>56.60</td>
<td>1.77</td>
<td>9.09</td>
<td>1.52</td>
</tr>
<tr>
<td>60 and over</td>
<td>70.45</td>
<td>2.20</td>
<td>7.69</td>
<td>1.28</td>
</tr>
<tr>
<td>All aged 16 and over</td>
<td>54.9</td>
<td>1.72</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Trend growth rate: all aged 16 and over</td>
<td>3.50</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WOMEN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 - 24</td>
<td>39.02</td>
<td>1.22</td>
<td>20.00</td>
<td>3.33</td>
</tr>
<tr>
<td>25 - 34</td>
<td>44.68</td>
<td>1.40</td>
<td>13.04</td>
<td>2.17</td>
</tr>
<tr>
<td>35 - 49</td>
<td>48.98</td>
<td>1.53</td>
<td>13.04</td>
<td>2.17</td>
</tr>
<tr>
<td>50 - 59</td>
<td>54.17</td>
<td>1.69</td>
<td>9.52</td>
<td>1.59</td>
</tr>
<tr>
<td>60 and over</td>
<td>53.85</td>
<td>1.68</td>
<td>16.67</td>
<td>2.78</td>
</tr>
<tr>
<td>All aged 16 and over</td>
<td>48.78</td>
<td>1.52</td>
<td>15.00</td>
<td>2.50</td>
</tr>
<tr>
<td>Trend growth rate: all aged 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source(s): Office for National Statistics, 2013b
Note(s): 1. Weighted data are only available for 2000 < t
2. Data on cigarette use were collected on a 2Y-basis for t < 2000
3. Estimate t < 2005 are based on a fiscal Y rather than a calendar Y
4. Estimates for 2004 and 2005 both include the period 01/01/05 to 31/03/05 due to a change in a survey period from a fiscal Y to a calendar Y
5. Estimates for 2006 to 2011 include longitudinal data
6. Weighted bases are given to the nearest thousand
From when the smoke-free legislation came into place, the effect on female prevalence rates has been greater than that of males, showing a 15% overall decrease for all females aged over 16 and a 2.5% annual average decrease. The largest decrease is amongst the youngest female cohort 16-24, where the prevalence level decreased at an increasing rate with the average annual decline going from 1.22% to 3.33%. This same age group amongst males showed the smallest change since the smoking ban. Despite smoking prevalence declining within this age group across both genders, it is still the cohort with the highest percentage of smokers. Consequently, with the time-lag effect, a burden will be placed on the NHS, as persons within this age group are the most susceptible to becoming addicted to nicotine (Department of Health, 2011). When comparing both male and female trend growth rates, the smoking ban has had a much greater effect on females, as prevalence rates have fallen by 2.6% compared to the male equivalent of 0.2%, despite the pre-ban trend growth rate being similar across both genders. This indicates that the female consensus about smoking is more uniform across all ages, and it may be inferred that the decline in smoking among females will fall further in years to come as fewer individuals resist inaugurating the addictive habit.

Figure 4.2: Proportion of Women Who Smoke Cigarettes in Great Britain, 1974-2013

Through running a paired t-test, the null hypothesis, i.e. that the mean smoking prevalence levels of men is equal to women, can be rejected. The paired t-test results show that men have a higher smoking prevalence (32.90 ± 7.80) compared to females (29.08 ± 5.59). Furthermore, pre smoking
A t-value of 6.7912 with 19 degrees of freedom (df) is given at a 95% confidence interval, resulting in a two-tailed p-value of 0.0001. As it is less than 0.05 (0<p<0.05), the results are statistically significant allowing the null hypothesis to be rejected. Similarly, post-smoking ban, the male smoking prevalence (21.62 ± 0.52) is higher than that of women (19.26 ± 0.48) and a t-value of 4.6588 with 6 df at a 95% confidence level, the two-tailed p-value (0.0035) is less than 0.05 (0<p<0.05) meaning the results are statistically significantly different, thus also rejecting the null hypothesis post-smoking ban.

Figure 4.3: Proportion Who Smoked Cigarettes, Proportion of Smokers Who Have Quit and Proportion Who have never Smoked, 1974-2013

Source(s): Office for National Statistics, 2013b
Note(s): 1. Weighted data are only available for 2000 < t
2. Data on cigarette use were collected on a 2Y-basis for t < 2000
3. Estimate t < 2005 are based on a fiscal Y rather than a calendar Y
4. Estimates for 2004 and 2005 both include the period 01/01/05 to 31/03/05 due to a change in a survey period from a fiscal Y to a calendar Y
5. Estimates for 2006 to 2011 include longitudinal data
6. Weighted bases are given to the nearest thousand

Figure 4.3 gives a summary of the proportions of the population in relation to smoking prevalence in three different categories: proportion that smokes, proportion that has quit and proportion that has never smoked. Post 1982, it is clear to see that both the proportion of smokers who have quit and the
proportion of those who have never smoked cigarettes are substantially higher than the proportion of those who smoke cigarettes. This coincides with the negative health effects of smoking being publicised; making the public more aware of the health burden smoking creates. Between 1974 and 2013, the overall decline in the proportion that smoke cigarettes has fallen by 58.7%, and the percentage of those that have never smoked has increased by 56.76% in the same time period. After the smoking ban, the percentage of persons who have never smoked has increased by 5.45% and the percentage of those that smoke has fallen by 4.76% over the last 5 years shown in Figure 4.3. Contrary to what has been found earlier, it could be inferred that although the relative percentage changes are small, there has been a positive effect on smoking prevalence since the smoking ban came into place, as a higher proportion of the population ceases smoking or never starts in the first place. Since the smoking ban came into place, there is an upward trend in the proportion of the population who never start smoking, which may now only be realised due to the time lag effect and younger generations being the most influenced. Both De Chaisemartin et al. (2011) and Anger et al. (2011) have found that although smoking per se has not declined throughout the whole population with the introduction of the smoking ban, there has been an increased demand for smoking cessation services among heavy smokers and consumption rates amongst ‘social smokers’ has decreased.

4.2 Statistics on the Proportion of Smokers by Socio-Economic Classification and Smoking Status

The socio-economic variable, ‘occupation’, discussed in the literature review, is further explored here to help quantify whether occupation influences smoking prevalence. Table 4.2 shows smoking status by occupation and Figure 4.4 shows the proportion of smokers who have quit by occupation. ‘Managerial and professional occupations’ are thought to be those that often work in large organisations, have a professional qualification or are self-employed, ‘intermediate occupations’ are those primarily in services, sales or clerical, whereas ‘routine and manual occupations’ consist mainly of tradesmen. It is clear to see that the lowest occupational group of ‘routine and manual occupations’ has a significantly larger percentage of smokers, at 29%, compared to the highest occupational group ‘managerial and professional occupations’ at 14%. Evidently, the number of persons that have never smoked in the highest occupational group is considerably greater than that of the other two occupations. This may be due to social norms within each classification, as prevalence levels are highest amongst lower occupational groups making it more acceptable to smoke with your peers. Graham et al. (1999) revealed that there was a strong socio-economic correlation with consumption of tobacco; smokers with educational qualifications had a mean consumption of almost three cigarettes less than those with no qualifications. To further support this, an OPCS study found that the proportion of women who smoke more than 20 cigarettes a day changes from 41% in the lowest socio-
economic group to 23% in the highest (Office of Population Census and Surveys, 1996), thus highlighting differences in socio-economic norms. However, it is interesting to note the uniformity in the percentage of ‘ex-smokers’ across all three occupational statuses, not deviating from 22-23%. A prominent factor contributing to socio-economic inequalities is the percentage of those that have ‘never smoked’ within each classification.

Table 4.3: Cigarette Smoking Status by Occupational Classification, Great Britain, 2013

<table>
<thead>
<tr>
<th>Cigarette Smoking Status by Occupational Classification, Great Britain, 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Managerial and professional occupations (%)</td>
</tr>
<tr>
<td>Smoker</td>
</tr>
<tr>
<td>Ex-smoker</td>
</tr>
<tr>
<td>Never smoked</td>
</tr>
</tbody>
</table>

Weighted base (000s)

| | Managerial and professional occupations | Intermediate occupations | Routine and manual occupations |
|---|---|---|
| Smoker | 14,248 | 8,505 | 11,910 |
| Ex-smoker | 3,330 | 2,000 | 2,800 |

Source(s): Office for National Statistics, 2013b
Note(s): 1. The group ‘ex-smoker’ contains those who said that they do not smoke cigarettes nowadays, but who have smoked cigarettes regularly in the past
2. The group ‘never smoked’ contains people who said that they do not smoke cigarettes nowadays, and those that have never smoked cigarettes regularly
3. Weighted biases are given to the nearest thousand

Figure 4.4 looks further into the proportion of smokers who have quit by occupational classification. It is important to note that Figure 4.4 is representative of the proportion of all ex-smokers subcategorised into occupational classification, whereas Table 4.2 is occupational classification subcategorised into proportions of current smoking status, hence the different percentages shown. Figure 4.4 shows that ‘managerial and professional’ occupations have the highest quit rate at 38%, which is significantly higher than those in ‘routine and manual’ occupations. It has been shown that within Britain between the 1970’s and 1990’s, the cessation rate among those in higher socio-economic class almost doubles, whereas amongst the poorest groups cessation rates have changed very little (Graham et al., 1999).
4.4 Statistics on the Perceived Success of the Smoking Ban

In 2013, the HSE produced data on the attitudes to smoking. When 1,089 smokers were presented with the theoretical question ‘what are your main reasons for wanting to give up smoking?’ out of the nine possible answers\(^1\), Figure 4.5 graphically represents only four of these, as they are the most relevant to this paper. It becomes apparent that the smoking ban was the least motivational reason for smokers wanting to give up, with 1,058 people voting ‘no’ as their answer. The most popular

\(^1\) The nine possible answers were: Due to a current health problem? For better health in general? For less risk of smoking related illnesses? Due to the ban on smoking in public places? Due to family/friends? Financial main reason? Worried about the effect on children? Worries about effect on other family members? Other? (Health Survey for England, 2007)
reason for wanting to give up smoking was for ‘better health in general’, with 783 people stating this aspect as being the primary reason. The reason ‘less risk of smoking related illnesses’ was the most evenly split question, resulting in 646 people answering ‘no’ and 443 answering ‘yes’. This shows that the impact of the smoking ban was minimal in encouraging cessation amongst current smokers in comparison to improving general health. These findings are consistent with Jones et al. (2015, p.175) as they also conclude that ‘the introduction of the public smoking bans in England and Scotland had limited short-run effects on both smoking prevalence and the total level of smoking’. Importantly, emphasis on ‘the short run’ must be placed, as benefits from cessation may not be fully realised, thus highlighting the time-lag effect once again.

Figure 4.5: What are your main reasons for wanting to give up smoking?

Figure 4.6 gives an indication on individuals’ perceptions about the success of the smoking ban the year it was brought in in 2007. The valid cases were asked various questions pre-smoking ban and a follow up of these questions was asked post-smoking ban, all of which are presented in Figure 4.6. As is evident from the results, the smoking ban hasn’t caused more people to stay at home and smoke, neither has it caused the majority of regular smokers to cut back the number of cigarettes smoked. Despite smokers being more inclined to smoke at home and in the car, it has been shown that there
have been fewer incidents of children being exposed to smoke at home (Department of Health, 2011), thus helping to reduce the number of SHS hospital admissions. It is important to recognise that the smoking ban in pubs has had a higher impact than in restaurants. These self-reported changes in tobacco consumption pre- and post-smoking ban, alongside the reduction in SHS, may have influenced the number of smoking related hospital admissions. Section 4.4 looks further into this.
Figure 4.6: Individuals perceptions on whether the smoking ban has been successful on cutting back the number of cigarettes they smoke, and whether it has been successful, 2007

Source: Health Survey for England, 2007
Analysing the number of NHS hospital admissions attributable to smoking from 1996/97 to 2011/12, will help provide further analysis on the success of the smoking ban. Figure 4.7 shows that despite smoking prevalence levels decreasing within this time space, there has been a 40% increase in the number of hospital admissions caused by smoking in adults aged 30 and over. Even when taking into account the changing population in Britain, that is 58.17 million mid-1996 (Office for National Statistics, 2004) and 63.3 million in mid-2011 (Office for National Statistics, 2013a), the number of hospital admissions has risen from 1.93% to 2.47%. This could be due to the time-lag effect, and the health benefits from the decrease in prevalence levels hasn’t been fully realised yet as smokers predating 2007 still need substantial healthcare. Allender et al. (2009) also encountered similar limitations, claiming that it is difficult to compare healthcare cost estimates year on year, as methodological differences such as changes in resident population prohibit easy comparison. Only ages 35 and over are presented in Figure 4.7, as this is the age where most damaging health effects are realised.

Table 4.3 shows in greater depth the two main types of disease attributable to smoking at primary diagnosis: cancers and respiratory diseases. The number of cancers is consistently higher than the number of respiratory disease across all the years shown in Table 4.3. However, of the total number of cancer deaths and respiratory deaths, the percentage attributable to smoking is smaller among cancer deaths, reaching 29% in 2009, compared to 35% for respiratory diseases (Department of Health, 2011). Cancers attributable to smoking have increased by 50.28% from 1996 to 2012 whereas the respective increase in respiratory diseases is 80.25%. However, since the smoking ban came into place, the number of cancers attributable to smoking is increasing at a substantially decreasing rate; represented by the trend growth falling from 2.9% to 0.7%. This may provide evidence that the smoking ban, alongside other tobacco control policies, has been successful in reducing the number of hospital admissions attributable to smoking. Contrary to that, the number of respiratory diseases is increasing at an increasing rate, shown by the trend growth going from 3.7% pre-ban to 6.1% post-ban. This may be due to the time lag from when the smoking ban came into effect or to the advancement in technology in detecting such diseases. Despite the number of cancers attributable to smoking being higher than that of respiratory disease, the overall treatment cost is substantially higher for the latter – amounting to 24% compared to 16% for cancer (Callum et al., 2011). The hospital admissions represented in Table 4.3 includes those affected by SHS. There is sufficient evidence to show that involuntary smoking, through SHS, is a direct cause for lung cancer in humans (International Agency for Research on Cancer, 2004), thus affecting the number of hospital admissions.

4.4 Statistics on NHS Hospital Admissions

Table 4.3 shows in greater depth the two main types of disease attributable to smoking at primary diagnosis: cancers and respiratory diseases. The number of cancers is consistently higher than the number of respiratory disease across all the years shown in Table 4.3. However, of the total number of cancer deaths and respiratory deaths, the percentage attributable to smoking is smaller among cancer deaths, reaching 29% in 2009, compared to 35% for respiratory diseases (Department of Health, 2011). Cancers attributable to smoking have increased by 50.28% from 1996 to 2012 whereas the respective increase in respiratory diseases is 80.25%. However, since the smoking ban came into place, the number of cancers attributable to smoking is increasing at a substantially decreasing rate; represented by the trend growth falling from 2.9% to 0.7%. This may provide evidence that the smoking ban, alongside other tobacco control policies, has been successful in reducing the number of hospital admissions attributable to smoking. Contrary to that, the number of respiratory diseases is increasing at an increasing rate, shown by the trend growth going from 3.7% pre-ban to 6.1% post-ban. This may be due to the time lag from when the smoking ban came into effect or to the advancement in technology in detecting such diseases. Despite the number of cancers attributable to smoking being higher than that of respiratory disease, the overall treatment cost is substantially higher for the latter – amounting to 24% compared to 16% for cancer (Callum et al., 2011). The hospital admissions represented in Table 4.3 includes those affected by SHS. There is sufficient evidence to show that involuntary smoking, through SHS, is a direct cause for lung cancer in humans (International Agency for Research on Cancer, 2004), thus affecting the number of hospital admissions.
admissions. The smoking ban has had a positive impact on the levels of passive smoking, yet in 2008 lung cancer was the third most common type of cancer among women (Cancer Research UK, 2011). In regard to hospital admissions it is important to consider how the population has increased between the years 1996/97 to 2011/12 from 58.17 million (Office for National Statistics, 2004) to 63.3 million (Office for National Statistics, 2013a) respectively. Consequentially, these figures indicate that smoking, relative to population increases, has caused a smaller percentage of hospital admissions.

**Figure 4.7:** NHS Hospital Admissions for Adults Aged 35 and over, where there was a Primary Diagnosis of Diseases, which can be Caused by Smoking in Great Britain, 1996/97 – 2011/12

Source(s): Lifestyle Statistics, Health and Social Care Information Centre, 2013
Note(s): 1. The data include private patients in NHS Hospitals (but not private patients in private hospitals)
2. Figures are presented for adults aged 35 years and over except for admissions Age related cataracts where patients must be 45 years and over and admission for hip fracture where patients must be aged 55 years and older due to risk rations only being available for these age groups
3. The figure excludes people whose gender was unknown or unspecified and whose country of residence was not confirmed as England
4. The primary diagnosis is the first of up to 20 (14 from 2002-03 to 2006-07 and 7 prior to 2002-03) diagnosis fields in the Hospital Episode Statistics (HES) data set and provides the main reason why the patient was admitted to hospital
5. Figures exclude admissions for cervical cancer whose gender was specified as male
Table 4.3: NHS Hospital Admissions for Adults Aged 35 and over for Cancers and Respiratory Diseases Caused by Smoking in Great Britain, 1996/97 – 2011/12

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of cancers which can be caused by smoking</th>
<th>Number of Respiratory diseases which can be caused by smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996/97</td>
<td>224432</td>
<td>142268</td>
</tr>
<tr>
<td>1997/98</td>
<td>253268</td>
<td>139481</td>
</tr>
<tr>
<td>1998/99</td>
<td>265331</td>
<td>163532</td>
</tr>
<tr>
<td>1999/00</td>
<td>276897</td>
<td>166146</td>
</tr>
<tr>
<td>2000/01</td>
<td>274216</td>
<td>152154</td>
</tr>
<tr>
<td>2001/02</td>
<td>273228</td>
<td>161897</td>
</tr>
<tr>
<td>2002/03</td>
<td>283503</td>
<td>168838</td>
</tr>
<tr>
<td>2003/04</td>
<td>287919</td>
<td>189903</td>
</tr>
<tr>
<td>2004/05</td>
<td>294443</td>
<td>195817</td>
</tr>
<tr>
<td>2005/06</td>
<td>317774</td>
<td>197908</td>
</tr>
<tr>
<td>2006/07</td>
<td>324936</td>
<td>201578</td>
</tr>
<tr>
<td>Pre Ban Trend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Growth (%)</td>
<td>2.9</td>
<td>3.7</td>
</tr>
<tr>
<td>2007/08</td>
<td>322570</td>
<td>203693</td>
</tr>
<tr>
<td>2008/09</td>
<td>332229</td>
<td>232078</td>
</tr>
<tr>
<td>2009/10</td>
<td>330513</td>
<td>231384</td>
</tr>
<tr>
<td>2010/11</td>
<td>327447</td>
<td>260819</td>
</tr>
<tr>
<td>2011/12</td>
<td>337291</td>
<td>259286</td>
</tr>
<tr>
<td>Post Ban Trend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Growth (%)</td>
<td>0.7</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Source(s): Lifestyle Statistics, Health and Social Care Information Centre, 2013

Note(s): 1. The data include private patients in NHS Hospitals (but not private patients in private hospitals)
2. Figures are presented for adults aged 35 years and over except for admissions for age related cataracts where patients must be 45 years and over and admissions for hip fracture where patients must be aged 55 years and older due to risk ratios only being available for these age groups
3. The figure exclude people whose gender was unknown or unspecified and whose country of residence was not confirmed as England
4. The primary diagnosis is the first of up to 20 (14 from 2002-03 to 2006-07 and 7 prior to 2002-03) diagnosis fields in the Hospital Episode Statistics (HES) data set and provides the main reason why the patient was admitted to hospital
5. Figures exclude admissions for cervical cancer whose gender was specified as male

4.5 Analysis on Primary Diagnosis and Cigarette Consumption

To aid comparing levels of smoking prevalence to hospital admissions, Figure 4.8 shows the correlation between these two aspects in person’s aged 35 and over.
Figure 4.8: Number of Primary Diagnosis, which can be caused by Smoking and the Proportion of cigarette smokers aged 35 and over, Great Britain, 1996-2012

Figure 4.8 uses the same data as Figure 4.7 alongside statistics on prevalence levels from ONS. This graph demonstrates that the proportion of the population, above 35 years of age, who smoke cigarettes, has decreased by almost 10% over the last 16 years and yet the number of cases of primary diagnosis attributed to smoking has risen by 39%, thus regardless of the number of smokers falling, the cost the NHS incurs is yet to diminish. This could be due to more advanced technology being available to treat and detect tobacco related diseases. Allender et al. (2009) also found that the
prevalence of smoking is decreasing, though less in recent years; total health care costs are increasing despite fewer persons smoking, though. This is attributed to the time-lag effect from the smoking ban being implemented. Studies dedicated to finding the NHS cost of smoking have realised that direct comparisons from year to year are increasingly hard to estimate as price and type of treatment change, alongside incidence levels, all of which have to be accounted for in the allocated NHS budget for smoking (Scarborough et al., 2011). The minimal decrease, of 2%, in the proportion that smoke cigarettes can be partially explained by the time lag-effect. In this instance the survey was taken five years post-smoking ban and smokers within the age category, of 35 years and over, are most likely to be already addicted and therefore less likely to change their smoking habits regardless of the smoke-free legislation. A majority of current smokers become addicted before turning 18 (Lifestyle Statistics, Health and Social Care Information Centre, 2013), thus, indicating that current smokers presented here have been addicted for over a decade and are more likely to be a burden to NHS services.

Section 4.6: Summary

Overall, there has been a significant decrease in smoking prevalence since 1974. Nowadays, amongst the population, it is more common not to smoke than to smoke. This indicates that policy implementation in general has worked; it appears, however, that the smoke-free legislation itself has had minimal impact. Following 2007 there has been a negligible impact on smoking prevalence, particularly in males, as there has been a 0% change in overall cigarette consumption post-smoking ban. Figure 4.5 gave further indication that the smoking ban had little or no impact in causing cessation amongst current smokers. Despite smoking prevalence levels decreasing, the level of hospital admissions attributable to smoking is yet to decline. This could be due to the time-lag effect, as the health benefits from the overall decreasing level of prevalence haven’t yet been realised. However, the NHS is currently experiencing the health repercussions from those that previously or currently smoked. This may allow the conclusion to be drawn that in the near future, prevalence levels will decrease further causing a decrease in the number of hospital admissions.

5. Conclusion

In analysing the time-series data presented in Section 4, it can be concluded that the smoking ban has had a minimal to no effect on overall prevalence of smoking. The smoke-free legislation has had a differential affect dependent on various cohorts. Females have been more sensitive to the changes in smoking legislation, specifically those aged between 16 and 24, whereas this same cohort amongst males shows an insignificant decline in smoking prevalence. In an occupational classification context, the smoking ban appears to have had more of an impact on those in a higher classification. Those in
‘managerial and professional’ occupations appear to be more inclined to stop smoking compared to those in ‘routine and manual occupations’. Overall there has been a substantial decline in smoking prevalence since 1974, however, the contribution of the smoke-free legislation appears to be minimal with the proportion of the population who still smoke being largely unchanged since 2007.

From studying smokers’ perceptions on the introduction of the smoking ban, it becomes apparent that the smoke-free legislation is an insignificant reason for current smokers to want to quit. A majority of current smokers said the primary reason they would wish to quit is for improvements in general health, not just a reduction in risk of tobacco related diseases. From the survey conducted pre- and post- smoking ban, it also becomes evident that smokers’ intentions to smoke less are significantly different to their actions to do so. This allows for the conclusion that the smoking ban has had little or no causal affect in decreasing consumption of tobacco.

In terms of the number of hospital admissions, it is conclusive that despite the smoking prevalence levels falling since 1996, the number of hospital admissions related to smoking is yet to decline. However, since the smoking ban came into place in 2007 the rate of hospital admissions amongst ages 35 and over is increasing at a decreasing rate, showing potential delayed effects of the ban. This could be due to a time lag, and the health benefits from the cessation of smoking will not be realised until further research is carried out in future years.

Due to limitations of the data, further analysis and research is needed to draw more satisfactory conclusions. There is limited data available for the cost in healthcare attributable to smoking-related diseases, thus, making it difficult to draw accurate conclusions about the cost to the NHS. Every patient has different treatment plans and healthcare support, thus, making it inaccurate to generalise a treatment cost of a tobacco related disease. A further limitation in this study is the various methods in accounting for whether smoking is the cause for disease or merely exacerbates it, demonstrating the complex relationship between factors affecting individuals’ health. Additionally, future research could be carried out into regional and ethnical cohorts, giving the study a geographical and cultural dynamic. Finally, it would be beneficial to carry out this study amongst future generation in order to reduce the time-lag effect from the smoking ban.

This study adds to current and past literature by collating statistics dating back as far as 1974, with a focus on analysing smoking trends pre- and post-smoking ban within the UK. This study provides a foundation on which cohorts have the highest smoking prevalence levels, thus, indicating which
population sub-groups tobacco control programs should target in the future to reduce overall tobacco consumption. This study attempts to isolate the impact the smoking ban has had on current smokers by asking their reasons behind wanting to quit, something which current literature has previously linked in with other tobacco control measures. Finally, this study correlates the changes in smoking prevalence levels to the number of tobacco related hospital admissions and in turn NHS tobacco related expenditure.
Bibliography


Callinan J.E., A. Clarke, K. Doherty, and C. Kelleher. 2010. Legislative smoking bans for reducing second-hand smoke exposure, smoking prevalence and tobacco consumption. *Cochrane Database of Systematic Reviews* 2010. 4. [no pagination].


The Impacts of Immigration on Wages and Unemployment in England: An Empirical Investigation

George Wibberley

The economic effects of immigration have become increasingly important within the UK over recent decades but there is no general consensus on the impact of immigration on unemployment and wages. Within the existing literature, a number of authors have attempted to empirically measure the impacts of immigration on labour market outcomes within the UK at a national level. However, my study aims to use regional level data from England across the period 2005 to 2012 to analyse the impacts of immigration on unemployment and wages. My results show that immigration has no significant effects on either unemployment or wages at a regional level.
1. Introduction

The impact of immigration on a host country’s labour market has been a much-discussed issue, especially in recent decades. Although global migration levels have remained fairly constant over the past fifty years, the economic impact of migration has become of increasing importance politically to receiving countries (Castles et al., 2014, p. 1). Furthermore, despite a significant wealth of contradictory evidence, public perception of the impacts of immigration on labour market outcomes seems to be largely negative. This is especially the case for the United Kingdom, where the majority of respondents to a survey conducted by the European Social Survey (2002, p. 28) agreed that wage levels were reduced by immigration (Card et al., 2012). This is despite most studies finding that immigration has only had a small impact on wages and unemployment (Dustmann et al., 2003b; Manacorda et al., 2012; Ottaviano and Peri, 2012).

In order to analyse the impacts of migration, we must first consider how to define immigration. Definitions of who is considered to be a migrant vary greatly between sources and this can have significant consequences when analysing migration flows empirically (Anderson and Blinder, 2014). This empirical analysis will use data on long-term international migration. A long-term international migrant can be defined as an individual “who moves to a country other than that of his or her usual residence for a period of at least a year” with the destination country becoming their new country of residence (United Nations, 1998, p. 97).

This study aims to add to the literature by taking a fresh approach to empirically analysing the impacts of immigration on unemployment and wages. Previous studies have used theoretical models or census data to analyse labour market impacts of immigration. However, there is a gap within the literature in terms of analysing immigration impacts at a regional level. My aim is to use aggregate data on immigrant inflows to each region of England between 2005 and 2012 to explore the labour market impacts of immigration. Specifically, I aim to empirically investigate whether immigration has had any significant impacts on the dependent variables wages and unemployment by running two separate instrumental variable regression models.

The paper will be structured as follows. Firstly, a review of the current literature will be carried out, in order to draw comparisons from similar studies that have already been conducted in other contexts. Then, a theoretical framework will be established to allow me to formulate hypotheses for my own empirical work. Next, my methodology will be outlined, including presenting my regression models.
and explaining my choice of variables. Finally, my regression results will be analysed and limitations of my methodology will be outlined to show areas for future research.

My results suggest that immigration has had no significant impact on either wages or unemployment at a regional level, even after allowing for a number of potential problems with the model specification. This may be because immigration tends to have very localised impacts on labour market outcomes, with no systematic effects even at a regional level.

2. Literature Review

2.1 Early Evidence

The economic impact of immigration has been of growing importance to academics and politicians alike, especially due to the large increase in migration to developed countries since 1975 (Lowell, 2007). However, most of the early literature is centred around immigration into the USA, with a lack of available data preventing studies on immigration into the UK until recent years (Dustmann et al., 2003b).

Early theoretical research has proved useful in highlighting two of the recurring themes that much of the current literature has to consider: the negative impact of immigration on the earnings of low-income native workers (Reder, 1963) and the problem of the endogeneity of immigrant inflows, in which unemployment may depend on immigration, but immigration rates may also depend negatively on unemployment (Fleisher, 1963).

In the early 1980s, the first attempts were made to construct theoretical models to analyse the impact of immigration on the US labour market (Gerking and Mutti, 1980; Johnson, 1980). Both papers’ models suggest that immigration has a negative impact on the wage rate of low-skilled native workers, but a positive effect on the earnings of high-skilled natives. This is consistent with the findings of Briggs (1975), who found that the vast majority of immigrants entering the US were competing with natives for low-skilled, low-paid jobs with high employee turnover rates.

Grossman (1982) highlights an important factor for further research, the extent to which native labour and immigrant labour can be seen as substitute inputs. If native and immigrant labour are seen as substitutes, then an increase in immigration will have a negative impact on the employment rates and wages of native workers. However, although Grossman (1982) finds that immigrant and native labour
are substitutes for each other, the impact of immigration on native wages and short run employment rates is fairly small.

2.2 First Empirical Work

Most of the early research into the impacts of immigration on the labour market had been largely theoretical and there had been little empirical work done to determine the impact of immigration on the U.S. labour market (Borjas, 1987). Although basic empirical approaches had been used to analyse the relationship between native and foreign-born workers (Borjas, 1986; Grossman, 1982), these studies had limited value for determining the extent of substitutability between the two groups, because they aggregated different immigrant groups into one population. Borjas (1987) provides one of the first empirical studies in which the impacts of immigrants on labour market competition are estimated where the immigrant population is disaggregated by country of origin. He agrees with Grossman (1982) that immigrant labour is a substitute for certain native demographics, but argues that immigrants are actually complements for other labour groups. Furthermore, a decrease in low-skilled native US workers’ earnings in the 1980s could be partly attributed to a weak negative correlation between the numbers of immigrants within a local labour market and the earnings of natives (Borjas, 1994).

More recently, Borjas (2003) has argued that workers with similar education standards but different experience levels are not perfect substitutes. As a result, he claims that immigration has lowered the wage of the average U.S native worker by 3 to 4 per cent, although the accuracy of these figures could be improved by reducing sampling errors in U.S. census data (Aydemir and Borjas, 2011).

Although Card (2001) also found that immigrant inflows had reduced the wages of low skilled workers in some U.S. cities, he argued that the overall impact of immigration on native wages and employment is negligible (Card, 2005). Similarly, Ottaviano and Peri (2012) have also challenged the previous assumptions about the impacts of immigration on wages. They suggest that immigration actually has a positive impact on the long run wages of the average native worker, with a slightly larger change in the wages of low-skilled native workers. Furthermore, the authors argue that the only negative labour market impact of immigration is a high negative effect (-6.7%) on the wages of previous immigrants (Ottaviano and Peri, 2012, p. 191).

Altonji and Card (1991) found that immigrant inflows have small and unsystematic effects on the employment levels of less-skilled native workers. The authors use an instrumental variable to correct
for the endogeneity of immigrant inflows, which increases the accuracy and reliability of their findings. Furthermore, immigrants and native workers can be seen as imperfect substitutes, which lowers the competitive pressure that immigrants place on the labour market and explains their lack of impact on employment levels (Ottaviano and Peri, 2008).

As discussed above, the evidence from the U.S. does not reach a consensus about the impact of immigration on wages, with estimates ranging from small positive to fairly significant negative effects. However, most of the literature agrees that the impact of immigration on unemployment in the U.S. is likely to be negligible, with the greatest impacts on low-skilled native employment levels (Altonji and Card, 1991; Ottaviano and Peri, 2008). This is supported by evidence from a number of other countries, which suggests that it is unlikely there is a negative long run impact of immigration on either wages or unemployment and any effects are usually fairly insignificant (Edo and Toubal, 2015; Friedberg, 2001; Friedberg and Hunt, 1995; Jean and Jiménez, 2011; Winter-Ebmer and Zweimuller, 1996)

2.3 UK Context

Research into the impacts of immigration on wages and unemployment in the UK has been a relatively recent phenomenon, with research only being conducted following the increase in immigration of working age adults in the late 1990s (Dustmann et al., 2003a, p. 19).

Dustmann et al. (2003b) conduct one of the first empirical studies into the impacts of immigration on the UK labour market, by analysing data across multiple sources. The authors use the instrumental variable of lagged immigrant concentrations to mitigate for the simultaneity problems of immigrants being attracted to areas with favourable labour market conditions. As was observed in the U.S. literature, immigration to the UK has a small, statistically insignificant impact on unemployment (Dustmann et al., 2003b). Furthermore, immigration appears to be positively associated with higher wage growth for native workers, although these findings may not be entirely accurate due to limitations in the available data and conceptual problems during the empirical analysis (Dustmann et al., 2003b). Similarly, Dustmann et al. (2005) analyse the impacts of immigration on UK wages and unemployment through looking at differences in immigration to various geographic areas, instrumented by variation in historical immigration patterns. The authors again find little evidence of adverse labour market effects arising from immigration at an aggregate level, although there may be negative effects for the employment of native workers with an intermediate education level (Dustmann et al., 2005). Indeed, a later study by Dustman et al. (2008a) also observes that immigration
has a positive effect on average native wages, with a negative impact on the lower end of the native wage distribution, but a positive impact on the higher end of the distribution.

Manacorda et al. (2012) adopt an approach similar to Ottaviano and Peri’s (2008) by suggesting that immigrants and native workers are imperfect substitutes in the UK labour market. As a result, the authors suggest that the only significant negative impact of increased immigration is to lower the wages of immigrants already in the country.

In 2004, the UK granted free movement to eight Eastern and Central European countries (A8 nations) that had joined the European Union, leading to a significant rise in immigration (Blanchflower et al., 2007). The potential negative impacts of increased immigration from A8 countries on the UK labour market has been a much discussed phenomenon (Lemos and Portes, 2008). However, there has been little empirical evidence of any negative impacts associated with A8 migration, although Portes and French (2005) reported small sectoral increases in agricultural claimant unemployment associated with higher migration post accession. Despite this, a number of studies have found that immigration from A8 countries to the UK has had no statistically significant impacts on the unemployment rate or wage level of any native workers (Gilpin et al., 2006; Lemos and Portes, 2008).

Furthermore, it has been proposed that A8 migration has actually helped to reduce the NAIRU through increasing aggregate supply (Blanchflower et al., 2007). This is because migrant workers are considered to be more hardworking and productive than many native workers, meaning that immigrants generally have a low propensity to be unemployed (Dench et al., 2006). As a result of being more productive, migrant labour has the potential to lower the natural rate of unemployment through filling skills gaps in the UK labour market or reducing wage demands, as wage bargainers become more easily replaceable (Saleheen and Shadforth, 2006). This suggests that increased immigration may have a positive impact on long term unemployment rates, although there is little empirical evidence to suggest any statistically significant impacts as yet.

Very few studies have used regional aggregate data to analyse the impact of immigration on unemployment and wages. However, Hatton and Tani (2005) used regional level data to investigate the possibility that an influx of immigrants to a region causes an increase in inter-region migration amongst the non-immigrant population of the UK. The authors found that there were no statistically significant results to suggest that immigration caused negative displacement effects on native
populations, although the main focus of the paper was on migration flows not labour market outcomes.

As we have discussed, the literature from the UK suggests that immigration has no significant impacts on unemployment, but some impacts on the wage distribution and earnings of certain skill groups (Dustmann et al., 2008a). This study attempts to add to the literature by taking a different approach in analysing net immigration inflows to specific regions of the UK, by analysing regional data to see if immigration has any major effects on unemployment and wages.

3. Theoretical Framework

Many early studies have argued that migrant and native workers are two separate components of labour and as such are not perfectly substitutable (Grossman, 1982). However, this framework will take a similar approach to Dustmann et al. (2008b) in splitting up different groups of labour input by skill level instead of differentiating between native and immigrant labour.

For this analysis, I have constructed a simplified model, which assumes that an economy produces only one output good, using a mixture of the inputs labour and capital. Within labour, workers are either skilled or unskilled with both categories being made up from a combination of native and immigrant labour. For now the model assumes that immigrants and native workers are perfectly substitutable and that all labour is perfectly inelastic, meaning that workers will work at any wage rate. Finally, the model assumes that capital supply is perfectly elastic, with a fixed interest rate.

Firstly, we imagine a situation in which the labour market is in equilibrium, with equal levels of highly skilled and unskilled labour. All workers are employed at equilibrium wage rates, with higher wages for more highly skilled workers. If incoming migrants have different skill levels to the current population, then the skill composition of the economy will change and wages and employment will have to adjust in the short run to absorb the immigrant labour. It is important to note that immigration will only affect wages and unemployment if the influx of migrant labour changes the skill composition of the economy (Dustmann et al., 2008b). This can be illustrated by the following scenario.

Suppose an extreme case, where the entire incoming immigrant workforce is highly skilled. This will create a surplus of highly skilled labour, which allows firms to lower the wage rate they offer for highly skilled labour. This increases demand until the point where all highly skilled labour is employed,
including both the native and immigrant populations. Therefore, highly skilled natives will lose out
due to immigration in the form of lower wage rates.

This can be illustrated by figure 3.1, which concentrates on the labour market impacts for highly skilled
labour. The vertical axis displays wages, whilst the horizontal axis shows employment levels. The pre-
migration equilibrium is represented by point A where all native highly skilled workers are employed
(E) at wage rate \(w_1\). Immigration of highly skilled labour of magnitude M then occurs. Because supply
of unskilled labour remains constant and labour supply is perfectly inelastic, this drives wages down
the marginal product curve. This results in a new equilibrium at point B, with highly skilled wages at a
lower equilibrium level of \(w_2\).

**Figure 3.1: Impact of Migration on Highly Skilled Wage Rates** (adapted from Dustman *et al.*, 2008b,
p. 480)

![Figure 3.1: Impact of Migration on Highly Skilled Wage Rates](image)

However, the increase in highly skilled labour causes an increase in demand which creates a relative
shortage in unskilled labour. This shortage increases the wage rate for unskilled labour. This can be
represented in figure 3.1 by the triangle A-B-C which shows the additional surplus created by highly
skilled immigration, which is then absorbed by unskilled labour. This is because all highly skilled
workers work at a wage equal to the marginal product of the last immigrant (Dustmann *et al.*, 2008b).
Figure 3.2 shows the impacts of highly skilled immigration on the wages of unskilled labour. The vertical axis shows the wages of unskilled workers, whilst the horizontal axis shows unskilled employment levels. Before immigration, the equilibrium lies at point A, where all unskilled labour is fully employed (point E) at wage rate $w_1$. However, the migration of highly skilled labour (M) causes an increase in demand for unskilled labour, shown by the outward shift in the labour demand curve from $D_1$ to $D_2$. Because unskilled labour supply remains constant at point E, post-migration, the increase in demand drives the wage rate up for unskilled labour to point $w_2$. The new equilibrium is point B. The triangle A-B-C represents the wage gains for unskilled labour as a result of the surplus created by highly skilled immigrant labour (see triangle A-B-C in figure 3.1). Thus, from the influx of immigration, the wages of highly skilled labour increases, but the wages of unskilled labour falls. The overall effect on average wages of immigration is zero.

Figure 3.2: Impact of Highly Skilled Migration on Unskilled Wage Rates (adapted from Dustman et al., 2008b, p. 480)

The above implications are based on the assumption that capital supply is elastic. If we give up this assumption, then immigration will redistribute wages from highly skilled labour to unskilled labour, but also to capital. Therefore, average wages may actually decrease as a consequence of immigration under these circumstances. The impact on average wages depends on the rate at which capital can adjust to immigrant inflows, with faster adjustment rates associated with smaller wage effects (Dustmann et al., 2008b). However, the authors argue that there is currently a lack of empirical
evidence that links capital flows to immigration. Therefore, it is not unreasonable to retain the assumption of perfectly elastic capital supply for the purposes of this investigation.

Another assumption that was initially made in the model is that of completely inelastic labour supply. In the initial example above, full employment is retained after the immigration influx, because highly skilled native workers are still willing to work at the new reduced wage rate. However, in reality, labour supply is likely to be elastic to some degree, with some workers willing to choose unemployment rather than accept a lower wage rate. Thus, in our example, immigration may cause voluntary unemployment amongst some of the highly skilled native workers, who are unwilling to work at the new equilibrium wage rate \( w_2 \) in figure 3.1. Under these circumstances immigration may have negative employment effects.

A further assumption that has been challenged is that natives and immigrants are substitutes within a given skill group. If natives and immigrants are seen as imperfect substitutes, immigration will only significantly affect immigrants who are already in the economy (Manacorda et al., 2012; Ottaviano and Peri, 2012).

Therefore, the theoretical impacts of immigration on wages and unemployment are dependent on the key assumptions holding. Under this model, immigration only impacts wages and unemployment if the influx changes the skills makeup of the receiving economy. Furthermore, the impacts of immigration will be felt differently by native workers across the distribution of wages. In reality, the theoretical impacts of immigration are less easily discernible, especially because economies do not produce one single output good (Dustmann et al., 2008b). However, most labour economists prefer a model as described above because it offers a strong structure for empirical work compared to more complex models with multiple final goods (Gaston and Nelson, 2000). Nevertheless, it must be kept in mind when analysing empirical results that the model used in this study is a vast simplification.

4. Data and Methodology

4.1 The Dataset

The empirical investigation will involve the analysis of a panel data sample consisting of observations for each region of England across the sample period from 2005 to 2012. A region is defined as having the same boundaries as former government offices for the region (GORs) (Office for National Statistics, 2015c). There are nine regions in England that will be used in the empirical analysis. The
sample period 2005 to 2012 has been selected because it is large enough to give a significant number of observations for each variable. The dataset has been formulated from a variety of sources published by the Office for National Statistics (ONS) and the Department for Education. Table 4.1 shows the variables that will be used in the investigation and the summary statistics for each.

**Table 4.1: Variable Summary Statistics**

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Mean</th>
<th>Variable Name</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net migration inflows</td>
<td>22.938</td>
<td>Gross value added</td>
<td>96.876</td>
</tr>
<tr>
<td></td>
<td>(24.550)</td>
<td>(17.121)</td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>6.740</td>
<td>High qualifications (%)</td>
<td>28.593</td>
</tr>
<tr>
<td></td>
<td>(1.848)</td>
<td>(5.399)</td>
<td></td>
</tr>
<tr>
<td>Wages</td>
<td>508.568</td>
<td>House prices</td>
<td>216.435</td>
</tr>
<tr>
<td></td>
<td>(72.955)</td>
<td>(65.221)</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>68.382</td>
<td>Low pay (%)</td>
<td>1.104</td>
</tr>
<tr>
<td></td>
<td>(10.538)</td>
<td>(0.247)</td>
<td></td>
</tr>
<tr>
<td>No qualifications (%)</td>
<td>12.444</td>
<td>National unemployment</td>
<td>6.625</td>
</tr>
<tr>
<td></td>
<td>(2.779)</td>
<td>(1.369)</td>
<td></td>
</tr>
<tr>
<td>EU migration regulations</td>
<td>0.750</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.463)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>72</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard deviations reported in brackets. The variable 'net migration inflows' is reported in thousands of people. The variable 'house prices' is reported in thousands of pounds.

**4.2 Dependent Variables**

The study will analyse the impact of immigration on the dependent variable unemployment using the International Labour Organisation’s (ILO) measure of unemployment, with data sourced from the Labour Market Statistics Dataset (Office for National Statistics, 2015f and 2015g). The ILO measure of unemployment has been used for this analysis because it gives a more realistic representation of the true extent of unemployment, whereas the claimant count measure excludes certain individuals who are unemployed but ineligible to claim Job Seekers Allowance (Clegg, 2011). The dependent variable ‘wages’ is also included in the form of average gross weekly earnings of full-time employees by region.
This gives an indication of variations in general wage levels, which can easily be compared between regions.

4.3 Independent Variables

The main independent variable being examined in this empirical investigation is long-term international migration flows into the UK. Statistics on regional migration to the UK were produced as part of the ONS’s Local Area Migration Indicators (2014b). Net migration inflows were then calculated for each region from this data by subtracting long-term international migration outflows from long-term international migration inflows.

The analysis also includes a number of other explanatory variables that are likely to determine wages and unemployment. The purpose of this is to reduce the likelihood of omitted variable bias, which may impact the accuracy of any coefficients estimated by the model (Clarke, 2005). The variable Gross Value Added (GVA) has been included to give some indication of the economic output of each region of England, which is likely to be an explanatory factor for variations in wages and unemployment. The data was sourced from the labour productivity dataset (Office for National Statistics, 2015e) and gives the GVA per filled job for each region as an index, where the UK average is equal to 100. This allows for comparison between regions to assess which areas contribute more heavily to economic output.

The variable ‘education’ was included in the study as indicated by the number of pupils attaining 5 or more GCSEs or equivalent qualifications graded between A* and C (Department for Education, 2015). Regions with higher average levels of education may be more attractive to migrant labour and higher education may also be an explanatory variable for lower unemployment and higher wages within a region.

The explanatory variable of ‘low pay’ is included in the form of the percentage of jobs paid below the National Minimum Wage for each region (Office for National Statistics, 2014a). As many of the impacts of migration affect low-income workers more heavily, it is important to include an indication of the regional extent of low paid jobs.

It is also important to include variables that account for differences in the average qualification levels of employees within regions because this is likely to impact heavily on wages and unemployment. Therefore, the percentage of people aged 16 to 64 who have no qualifications and the percentage of
people aged 16 to 64 who have qualifications level NVQ4 and above (highly skilled) have been included as explanatory variables (Office for National Statistics, 2015a).

House prices within a region may also be an important factor to consider, because migrant labour may be attracted to areas where house prices are lowest. The average annual price of all dwellings within a region (Office for National Statistics, 2015d) is therefore included as an explanatory variable.

The national unemployment rate (Office for National Statistics, 2015g) has also been included as an explanatory variable, as this will be a major determinant of annual variations in regional unemployment rates. The average annual unemployment rate for each year within the sample has been repeated for each region within England.

4.4 Instrumental Variable

As discussed above, one of the major issues with this investigation is that the direction of causality between immigration and changes in labour market outcomes is not completely clear (Altonji and Card, 1991; Fleisher, 1963). This is because immigrants are likely to be attracted to more economically successful regions (Migration Advisory Committee, 2008, p. 38), meaning that immigrant inflows are dependent on wages and unemployment, as well as wages and unemployment being dependent on immigrant inflows (Dustmann et al., 2003b). As a result of this, any estimates of the impacts of immigration on unemployment and wages are likely to be biased.

In order to correct for this, an instrumental variable regression can be used. This involves finding an instrument that is correlated with the independent variable, but not correlated with the error term of the regression (Cameron and Trivedi, 2005, p. 97). For this investigation, an instrument would be a variable that is correlated with immigrant inflows but not related to any of the labour market outcomes. An instrumental variables regression can then be performed that uses the information the instrument provides to find unbiased estimates of the effects of immigration (Dustmann et al., 2003b).

This investigation uses the instrumental variable of changes to immigration legislation in the UK following the accession of Romania and Bulgaria (A2 nations) to the European Union in 2007. The legislation allowed increased access to the UK labour market for agricultural workers, highly skilled labour, students and self-employed workers from A2 nations (Migration Advisory Committee, 2008). This therefore represents a suitable instrumental variable, as it is an exogenous change that is likely to be positively correlated with immigrant inflows, but is unlikely to be correlated with labour market
outcomes. This is shown by a more than proportional increase in the number of A2 nationals residing in the UK from pre-accession to post-accession (Migration Advisory Committee, 2008, p. 61).

The exogenous change of immigration regulations in 2007 has been included as an instrument, in the form of a dummy variable where 0 represents pre-accession immigration regulations in 2005 and 2006 and 1 is equal to post-accession regulations for the remainder of the sample period. The inclusion of this instrumental variable should help to correct for biased estimates associated with the endogeneity problems described above.

4.5 The Model

For the empirical investigation, two separate regression models will be run which are as follows:

\[
(1) \text{Unemployment}_{it} = \beta_1 \text{Migration Inflows}_{it} + \beta_2 Wages_{it} + \beta_3 \text{Gross Value Added}_{it} + \beta_4 \text{Education}_{it} + \beta_5 \text{No Qualifications}_{it} + \beta_6 \text{Highly Qualified}_{it} + \beta_7 \text{House Prices}_{it} + \beta_8 \text{Low Pay}_{it} + \beta_9 \text{National Unemployment}_{it} + \beta_{10} \text{EU Migration Laws}_{it} + \mu
\]

\[
(2) \text{Wages}_{it} = \beta_1 \text{Migration Inflows}_{it} + \beta_2 \text{Unemployment}_{it} + \beta_3 \text{Gross Value Added}_{it} + \beta_4 \text{Education}_{it} + \beta_5 \text{No Qualifications}_{it} + \beta_6 \text{Highly Qualified}_{it} + \beta_7 \text{House Prices}_{it} + \beta_8 \text{Low Pay}_{it} + \beta_9 \text{National Unemployment}_{it} + \beta_{10} \text{EU Migration Laws}_{it} + \mu
\]

Where \(i\) denotes each region of England, \(t\) each year within the sample period, and \(\mu\) the error term.

All variables within the model are as defined above, where EU Migration Law is an instrument for migration inflows, in the form of a dummy variable as explained above. Both models follow a standard normal distribution.

Since the model uses panel data, ordinary least squares (OLS) regression estimators cannot be used. Instead, fixed effects models or random effects models are used to analyse panel data. This model will use a fixed effects estimator to remove the influence of common effects that may be present when assessing the impacts of immigration on labour market outcomes. Fixed effects may occur in the form of immigrant concentrations and labour market outcomes being spatially correlated due to historical settlement patterns or government policies (Dustmann et al., 2003b). This means that any positive or negative correlations that are found may not be attributed to genuine effects of immigration. The presence of these effects allows us to use the fixed effects estimator because the study aims to compute the magnitude of common effects within a population and not extrapolate those results to other populations (Borenstein et al., 2009).
As discussed in the literature review, there is no definite conclusion on the effects of immigration on wages and unemployment, although most authors agree that the impacts on unemployment are generally small in magnitude. I expect to find that immigration will not have a hugely significant impact on either wages or unemployment levels, although any effects on wages are likely to be positive.

5. Results

5.1 Unemployment

Table 5.1 shows the results of the regressions run using model 1 for the dependent variable unemployment. The table also shows the summary statistics, which indicate the reliability and usefulness of the model. As the model follows a standard normal distribution, the Z-values and P-values will be used to analyse the statistical significance of each coefficient, with the P-values reported in Table 5.1.

The R-squared value is a measure of the overall fit of the model and indicates the percentage of variance in the dependent variable that can be explained by the variance in the independent variables (Floyd, 2010, p. 203). The overall R-squared value for the unemployment model is 0.4908, which implies that the model has a fairly high goodness of fit. The value indicates that the independent variables are effective at explaining 49.08% of the overall variance in regional unemployment.

Furthermore, the F-statistic and associated probability indicate that at least one of the independent variables is a determinant of unemployment. The F-statistic is used to evaluate the null hypothesis that all the coefficients in the model are simultaneously equal to zero (Floyd, 2010, p. 239). The probability of obtaining an F-statistic test statistic that is greater than or equal to the null hypothesis value is 0.0230. This means that we can reject the null hypothesis at the 5% significance level and as such we can assume with a fairly high degree of certainty that at least one of the independent variables has a coefficient that is not equal to zero.

However, it is clear from the model that net migration inflows have no significant impact on unemployment levels when analysing regional level data. This is illustrated by the high P-value for the net migration inflow coefficient. The P-value indicates the support or lack of support provided by the sample for the null hypothesis that the coefficient is equal to zero (Anderson et al., 2010, p. 916). The P-value for net migration inflows is 0.724, which means that we cannot reject the null hypothesis that the net migration inflows coefficient is equal to zero. As such, we can deduce that the impact of
immigration on unemployment is small in magnitude and statistically insignificant. These results tend to support the findings of Dustmann et al. (2003b), Gilpin et al. (2006) and Lemos and Portes (2008), who also observed insignificant impacts of immigration on unemployment.

One potential explanation for this is that the vast majority of the variance in unemployment can be explained by regional variance. This is indicated by the fact that 89.8% of the variance within the model is regional variance, with none of the other independent variables accounting for a high proportion of the variance in unemployment. This would be in keeping with the findings of Dustmann et al. (2003b) who stated that variations in immigrant concentrations are very localised and often follow historic settlement patterns. As such, immigration effects on unemployment are likely to be insignificant on a regional level, with only certain very localised impacts occurring in smaller localities within regions.

Furthermore, the high P-values for all the other independent variables shows that none of the effects predicted are statistically significant at any level. This would reinforce the argument that much of the variance in unemployment is due to differences between regions as opposed to impacts of any of the independent variables over time.

The high magnitude of the standard error (0.265) of net migration inflows compared to the coefficient value (0.093) indicates the high degree of uncertainty of the impacts of immigration on unemployment predicted by the model. One possible explanation for this could be the relatively small sample size used to carry out the analysis, with only 72 observations for each variable. Nevertheless, it is clear from the model that there is no statistically significant impact of immigration on unemployment when analysing regional level data.

5.2 Wages

Table 5.2 shows the results of the regressions run for model 2 for the dependent variable ‘wages’. The table also shows the summary statistics for the model as well as an indication of the statistical significance of the results in terms of the standard errors and P-values for each variable.

The R-squared value for the wages model is 0.6186. This is higher than the value for the unemployment model and shows that the wages model has a relatively good level of fit. The value implies that the independent variables are effective at explaining 61.86% of the overall variance in regional wage levels.
However, the probability of obtaining an F-statistic test statistic that is greater than or equal to the null hypothesis value is 1.000. This means that we cannot reject the null hypothesis that all the coefficients in the model are equal to zero at any significance level. This implies that any one of the independent variables is not necessarily a significant determinant of wages.

Furthermore, it is clear from the model that net migration inflows has no statistically significant impact on wages. This is illustrated by the high P-value of 0.881 for the migration coefficient. This means that we cannot reject the null hypothesis that the migration coefficient is equal to zero at any significance level. There is therefore no statistically significant impact of immigration on wages at a regional level. These results tend to support the findings of Manacorda et al. (2012) and Ottaviano and Peri (2008) who also found that immigration had no adverse effects on wage rates of the population as a whole.

One of the main reasons for this could be the high fraction of the variance that is attributed to regional variations. The variable ‘region’ accounts for 97.5% of the total variance in wages, which indicates that none of the independent variables are likely to have significant effects on wages over time. This would tend to support the theory that immigration only has very localised effects on the wages of certain communities within regions, with no significant effects across regions as a whole.

As with the unemployment model, the high P-values for all the independent variables in the model indicate that none of them have a statistically significant impact on wages. This is likely to be because much of the variations in wages are due to specific regional factors instead of specific independent variables over time.

The net migration inflows coefficient has a high standard error of 213.475, which shows the high level of uncertainty regarding the coefficient predicted by the model. This could be explained by the relatively small sample size used in the analysis. Even accounting for this, it is clear that immigration has no statistically significant impacts on wages at a regional level.
**Table 5.1: Determinants of Regional Unemployment, 2005-2012**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Dependent Variable: Average Weekly Earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net Migration Inflows (Thousands)</td>
<td>0.093 (0.265)</td>
</tr>
<tr>
<td></td>
<td>[0.724]</td>
</tr>
<tr>
<td>Gross Value Added</td>
<td>0.016 (0.101)</td>
</tr>
<tr>
<td></td>
<td>[0.876]</td>
</tr>
<tr>
<td>Education (Attainment of 5 GCSEs A*-C)</td>
<td>0.120 (0.121)</td>
</tr>
<tr>
<td></td>
<td>[0.320]</td>
</tr>
<tr>
<td>No Qualifications</td>
<td>0.201 (0.493)</td>
</tr>
<tr>
<td></td>
<td>[0.684]</td>
</tr>
<tr>
<td>High Level of Qualifications</td>
<td>-0.001 (0.327)</td>
</tr>
<tr>
<td></td>
<td>[0.998]</td>
</tr>
<tr>
<td>Regional Weekly Earnings</td>
<td>-0.002 (0.017)</td>
</tr>
<tr>
<td></td>
<td>[0.923]</td>
</tr>
<tr>
<td>House Prices (Thousands of Pounds)</td>
<td>-0.002 (0.010)</td>
</tr>
<tr>
<td></td>
<td>[0.857]</td>
</tr>
<tr>
<td>Low Pay (below National Minimum Wage)</td>
<td>-0.036 (0.888)</td>
</tr>
<tr>
<td></td>
<td>[0.967]</td>
</tr>
<tr>
<td>National Unemployment Rate</td>
<td>0.505 (0.837)</td>
</tr>
<tr>
<td></td>
<td>[0.546]</td>
</tr>
<tr>
<td>Overall R- squared</td>
<td>0.4908</td>
</tr>
<tr>
<td>F-test Statistic.</td>
<td>2.48</td>
</tr>
<tr>
<td>Prob &gt; F</td>
<td>0.0230</td>
</tr>
<tr>
<td>Fraction of Variance due to Region</td>
<td>0.898</td>
</tr>
</tbody>
</table>

Note: Standard errors are reported in brackets. Two-tailed p-values are reported in parentheses.
<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Dependent Variable: Average Weekly Earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net Migration Inflows (Thousands)</td>
<td>32.042</td>
</tr>
<tr>
<td></td>
<td>(213.475)</td>
</tr>
<tr>
<td></td>
<td>[0.881]</td>
</tr>
<tr>
<td>Gross Value Added</td>
<td>12.610</td>
</tr>
<tr>
<td></td>
<td>(75.748)</td>
</tr>
<tr>
<td></td>
<td>[0.868]</td>
</tr>
<tr>
<td>Education (Attainment of 5 GCSEs A*-C)</td>
<td>17.400</td>
</tr>
<tr>
<td></td>
<td>(106.313)</td>
</tr>
<tr>
<td></td>
<td>[0.870]</td>
</tr>
<tr>
<td>No Qualifications</td>
<td>61.871</td>
</tr>
<tr>
<td></td>
<td>(400.116)</td>
</tr>
<tr>
<td></td>
<td>[0.877]</td>
</tr>
<tr>
<td>High Level of Qualifications</td>
<td>44.323</td>
</tr>
<tr>
<td></td>
<td>(281.966)</td>
</tr>
<tr>
<td></td>
<td>[0.875]</td>
</tr>
<tr>
<td>Regional Unemployment Rate</td>
<td>0.213</td>
</tr>
<tr>
<td></td>
<td>(69.755)</td>
</tr>
<tr>
<td></td>
<td>[0.998]</td>
</tr>
<tr>
<td>House Prices (Thousands of Pounds)</td>
<td>0.605</td>
</tr>
<tr>
<td></td>
<td>(2.607)</td>
</tr>
<tr>
<td></td>
<td>[0.816]</td>
</tr>
<tr>
<td>Low Pay (below National Minimum Wage)</td>
<td>3.195</td>
</tr>
<tr>
<td></td>
<td>(258.627)</td>
</tr>
<tr>
<td></td>
<td>[0.990]</td>
</tr>
<tr>
<td></td>
<td>-84.941</td>
</tr>
<tr>
<td>National Unemployment Rate</td>
<td>(621.686)</td>
</tr>
<tr>
<td></td>
<td>[0.891]</td>
</tr>
<tr>
<td>Overall R- squared</td>
<td>0.6186</td>
</tr>
<tr>
<td>F-test Statistic</td>
<td>0.02</td>
</tr>
<tr>
<td>Prob &gt; F</td>
<td>1</td>
</tr>
<tr>
<td>Fraction of Variance due to Region</td>
<td>0.975</td>
</tr>
</tbody>
</table>

Note: Standard errors are reported in brackets. Two-tailed p-values are reported in parentheses.

One potential explanation for the lack of significant results for both wages and unemployment is that immigration only has very localised labour market impacts. As such, empirical results for regions as a
whole are unlikely to be significant. This can be shown by the large differences in immigration flows of two more localised areas within a region. Figure 5.1 shows the net migration inflows for Coventry and Walsall over the sample period, two areas within the region of the West Midlands. For every year within the sample period net migration inflows were much higher in Coventry than in Walsall. Overall, Coventry saw an average annual inflow of 4587.5 migrants over the sample period compared to an average inflow of 487.5 migrants in Walsall. The large difference in the scale of immigration recorded by the two areas is especially significant because the two areas have similar population sizes, which are shown in Table 5.3. Furthermore, Coventry experienced much greater population growth over the sample period than Walsall, largely due to its increased levels of immigration.

**Figure 5.1: Net Migration Inflows for Coventry and Walsall, 2005-2012**

![Net Migration Inflows for Coventry and Walsall, 2005-2012](image)

Source(s): Net migration inflows were calculated from long-term international migration Flows produced by the Office for Nationals Statistics (2014b).
Table 5.3: Populations of Coventry and Walsall, 2005-2012

<table>
<thead>
<tr>
<th>Year</th>
<th>Population of Coventry (Thousands)</th>
<th>Population of Walsall (Thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>298.4</td>
<td>258.2</td>
</tr>
<tr>
<td>2006</td>
<td>300.1</td>
<td>259.5</td>
</tr>
<tr>
<td>2007</td>
<td>301.4</td>
<td>260.9</td>
</tr>
<tr>
<td>2008</td>
<td>305.2</td>
<td>263.0</td>
</tr>
<tr>
<td>2009</td>
<td>307.4</td>
<td>264.8</td>
</tr>
<tr>
<td>2010</td>
<td>311.7</td>
<td>266.8</td>
</tr>
<tr>
<td>2011</td>
<td>316.9</td>
<td>269.5</td>
</tr>
<tr>
<td>2012</td>
<td>323.1</td>
<td>270.9</td>
</tr>
</tbody>
</table>

Source: Annual mid-year population estimates were sourced from the Office for National Statistics (2014b).

Although this is an extreme example, it nevertheless illustrates the fact that migration is highly localised even within regions. This is largely due to the face that immigrants tend to be attracted to areas where there are already incumbent immigrant populations from the same country of origin (Dustmann et al., 2003b). Furthermore, factors that affect wages and unemployment are likely to be specific to certain more localised areas within regions, as indicated by the lack of significance of my results. As such, regional data does not allow us to fully analyse the labour market impacts of immigration and a greater disaggregation to smaller localities may yield more significant results (Hatton and Tani, 2005).

Overall the empirical work suggests that immigration does not have a significant impact on either wages or unemployment at a regional level. This reflects the predictions of the theoretical model outlined in chapter 3, in which immigration would only have significant effects on the labour market outcomes of certain skill groups within the economy.
6. Limitations

Despite controlling for many potential problems within my models, the empirical work conducted in this investigation has a number of limitations, which may detract from the conclusions that can be inferred.

One of the main limitations of the study is the potential presence of heteroscedasticity within the data. Heteroscedasticity occurs when the variance of the error term is not constant over the whole range of values (Floyd, 2010). This can cause an issue when creating regression models because heteroskedastic data can produce inefficient coefficient estimates (White, 1980). In terms of my model, heteroscedasticity may occur because the variance of the error term in the regression is unlikely to be constant across the entire sample period. This means that many of the tests for statistical significance may be invalid.

Furthermore, separating my model into two separate regressions for wages and unemployment respectively may have been an oversimplification. This is due to the fact that both nominal and real wages have a complex interdependent relationship with unemployment (Malinvaud, 1982). My unemployment model included wages as an independent variable to attempt to account for this relationship. Similarly, unemployment was included as an independent variable within the wages model. However, this is unlikely to have fully addressed the simultaneity problem between the two variables and as such this limits the conclusions that can be drawn from the estimates observed in this data.

Developing a model that uses a more complex method of estimation such as Hansen’s (1982) generalized method of moments (GMM) estimators may help to address some of these issues. Whilst GMM estimators can yield more efficient and consistent results, they led to lower R-squared values when used with my models and were therefore not used in my final results. However, future studies could work towards developing a more complicated model that uses GMM estimators to account for heteroscedasticity and simultaneity.

Whilst I attempted to reduce bias within my investigation, there may still be limitations within the data sources themselves. Observations for labour outcomes were sourced from the Office for National Statistics’ Annual Labour Force Surveys, which are based on surveys sent to a proportion of the UK population. As such, there is the potential for sampling bias within the dataset, as well as other limitations including a potential lack of adequate coverage of any industrial sector (Office for National
Statistics, 2011, p. 1). This might mean that the labour market for certain sectors of the economy are unrepresented, although the high sample sizes for the survey mean that overall bias is likely to be low.

Another potential limitation of the investigation is the suitability of the instrumental variable that was selected. The change in immigration legislation following the accession of the A2 nations to the European Union can be considered to be an instrumental variable because it is likely to be associated with slightly higher immigration levels. However, the magnitude of the increase in immigration associated with the legislation is very low due to the fact that the stock of immigrants from A2 nations represents only a small fraction of total immigration in the UK (Migration Advisory Committee, 2008, p. 60). Furthermore, because there were still heavy labour market restrictions on immigrants from A2 nations following the legislation change, there was not a vast increase in immigration following accession. However, the instrumental variable was selected because it represented the only significant change in immigration legislation over the chosen sample period.

Nevertheless, my estimates of the effects of immigration on wages and unemployment are likely to be slightly inaccurate due to the weakness of my instrumental variable. In order to improve my study, it would be better to analyse a different sample period in which there was a more significant change in immigration legislation that could be used as an instrumental variable. However, the most obvious immigration legislation change could not be selected as an instrumental variable due to problems sourcing regional level data. The introduction of the Worker’s Registration Scheme in 2004 following the accession of eight nations (A8) to the European Union was an important piece of legislation that led to increased immigration to the UK (McCollom et al., 2012). This legislation would have been a more effective instrumental variable to use because the scale of the increase in immigration following the A8 accession was much greater than the increase that followed the A2 accession. However, regional data for many of the independent variables that I used in my empirical work was not available before 2005 and this prevented the use of the A8 immigration legislation change.

Another potential choice of instrumental variable is the 2014 complete relaxation of labour market restrictions on A2 nationals, which gave Romanian and Bulgarian immigrants the same working rights as migrants from other EU member states (Gower and Hawkins, 2013). As a result of this legislation change, the Migration Advisory Committee (2008) predicted that immigration would increase significantly, which would therefore mean that the legislation change would make a suitable instrumental variable. However, regional immigration data is currently not available post 2013, which has prevented the use of the 2014 immigration legislation change as a variable for this investigation.

313
Nevertheless, when more years of data are available, this may prove to be a useful instrumental variable that could yield more significant results.

7. Conclusion

The aim of my investigation was to analyse the impacts of immigration on unemployment and wages using aggregate regional level data for all regions within England between 2005 and 2012. Firstly, this investigation assessed the relevant literature, before analysing the theoretical foundations that underpin the subject. This allowed me to construct a suitable methodology, which consisted of running two separate regressions for my dependent variables ‘wages’ and ‘unemployment’. I collected a dataset consisting of my two dependent variables, my main independent variable ‘net long-term international migration’ and seven other explanatory variables. For each variable, I used a panel dataset, which included observations for each of the nine regions of England for the years 2005 to 2012 inclusive. I decided to run my models as instrumental variable regressions in order to correct for endogeneity within my study.

My results suggest that there is no significant impact of immigration on either unemployment or wages at a regional level. This is largely because a high fraction of the variance in both dependent variables was due to regional factors. As a result, my migration coefficient and all of the other explanatory variables had no statistically significant impact on wages or unemployment. The results support the findings of Dustmann et al. (2003b) who also found that immigration has no impact on unemployment. The UK literature is divided on the impact of immigration on wage rates and my work tends to support the findings of Manacorda et al. (2012) and Gilpin et al. (2006) who found that immigration has negligible impacts on wages.

However, there are a number of limitations to my study that lower the significance of the conclusions that can be drawn from my findings. One of the main problems is that using regional data to analyse immigration impacts is ineffective because immigration is likely to be a much more localised phenomenon. Although my models attempted to correct for the presence of fixed effects and endogeneity, the highly complex relationship between the variables means that my models are still likely to be oversimplified. As a result, there may be some remaining inaccuracies within my estimations.
There is scope for my investigation to be continued using more recent data. This may allow for the selection of a more suitable instrumental variable that may fully correct for endogeneity in my model. Further investigation could also include analysis of the impact of immigration on wages and unemployment at a more disaggregated level by location, such as analysing outcomes within parliamentary constituencies. However, reliable data for all variables is not currently available.

Even allowing for the limitations within my model, it is hard to argue that immigration has any significant impact on either wages or unemployment at a regional level. This is most likely due to the fact that immigration tends to have specific impacts on smaller localities within regions. Despite contradictory empirical evidence that shows no negative impacts of immigration on unemployment and wages, the public perception of immigration appears to be relatively negative within the UK (German Marshall Fund, 2014). Therefore, policy makers need to do further research to determine whether increased immigration is beneficial or detrimental to labour market outcomes within the UK.
Bibliography


The Impacts of Digital Currency on China's Monetary System

Qianru Xiang

This dissertation contributes to the evaluation of how digital currency would affect the monetary system by studying the influence on money demand and supply, as well as the velocity of money. After exploring the characteristics of digital currency and comparing functions with fiat currency, this dissertation establishes an econometric model to explore further the relationship between digital currency, money supply and the velocity of money in the China monetary system. In order to quantify the effects of digital currency on the changing amount of fiat currency, by analysing cash ratio, electronic currency level, financial electronic level and interest rate, this research suggests that the extension of digital currency could bring variation to the velocity of money in both the long and short term. Although the initial stage of digital currency limits the results analysis, by discussing the impacts of digital currency on the central bank of China, this dissertation concludes with China’s monetary policy recommendations.
1. Introduction

1.1. Background of digital currency

1.1.1. Overview of digital currency
Since the 1970s, alongside the development of the electronic age and technical innovation, e-commerce has become a controversial topic throughout the world and during this time, an increasing number of Internet users has engaged in online business. Traditional cash payment is unlikely to catch up with the rapid expansion of the electronic commercial market. Therefore, as a complement to the fiat currency in use, emerging digital currencies have been continuously gaining popularity among the public. Starting in the 1990s, digital currency has been created for use with personal computers to meet the requirements of efficient e-commerce; it does not require complex legal licenses but enables online trading and has been gradually embraced by the market (Sloan, 2000; Plassaras, 2013). Digital currency has been recognised as a tool of a functioning economy, accompanying the progress of e-commerce in a growing number of business activities by individuals, firms and organisations, as well as becoming an official choice of nation states like Iceland (Smith and Weismann, 2014).

1.1.2. Significance of the Research in China
While digital currency has gained worldwide recognition, different countries employ digital currency to varying degrees. In China, the China Monetary Policy Report, published by the People’s Bank of China (PBOC) in 2013, provided official recognition that Internet finance can act as a complement to the financial system with a number of advantages, and believes it can enhance information transparency, lower transaction costs and achieve financial efficiency (PBOC, 2013a).

Digital currency is currently being embraced in China for an increasing number of Internet business activities. According to a report by the China Internet Network Information Center (2012), there are about 564 million Chinese Internet users, and approximately three-fifths of these Internet users have engaged in online payment. Consequentially, the value of e-commerce has reached over 800 million yuan, and online trading was over 1 trillion yuan in 2013 (China Economic News, 2013). Thereafter, in 2014, the online-trading value increased to 2.8 trillion yuan, with a rate of increase of 48.7% (Chinese Business Info, 2015).

Operating in a highly competitive environment, Chinese Internet companies have several products, such as Yu’ebao. Yu’ebao plays the role of an online wallet to store money and make transactions. It has similarities with the world-renowned payment system PayPal. Due to its minimal transaction fees,
Yu’ebao can provide an interest rate approximately five times higher than the rate offered by commercial banks (China Economy Network, 2013). In 2014, Yu’ebao had become China’s largest fund with assets of over 250 billion yuan (Jingu, 2014). Bitcoin is another novel form of digital currency with complete independence from the control of a central bank, the emergence of which has brought volatility to the market. Bitcoin allows direct payments between parties without involving third-party financial institutions or government authorities (Gloudeman, 2014). Since Bitcoin was created in 2009, China has provided more than half of the world’s Bitcoin trading, and above 100,000 Bitcoins, worth more than one hundred million dollars, were being traded daily in a Chinese Bitcoin exchange company, BTC China in 2013 (The Economist, 2014).

Those products have been claimed to contribute to the financial liberalisation driven by the market (Jingu, 2014). However, the circulation of digital currency in the market increasingly challenges the role of the Chinese central bank as the sole issuer of fiat currency (Qiu, 2013). Currently, the replacement effect of fiat currency with digital currency is becoming significant and an increasing number of researchers are beginning to focus on the impacts of the increasing use of digital currency.

1.2. Structure of the Research
1.2.1. Evolution of the Research into Digital Currency
In the early stages of digital currency, the majority of the literature focused on its characteristics, legalisation and risk management. Thereafter, along with the evolution of digital currency, the effects of replacing fiat currency with digital currency brought disputes to academic research.

Firstly, various potential explanations have been offered as to the explanation of the demand for digital currency. The most popular argument regarding the demand for digital currency is based on technological advances (Valdes-Benavides and Hernandez-Verme, 2014). Likewise, Casey and Vigna (2015) also associated the growth of digital currency with the advancement of technology, conversion and regulation. Riza (2010) suggested that, subject to market preferences, the amount of digital currency would continually increase in the coming years.

In addition to varying opinion as to the source of digital currency demand, another major debate regards the relationship between digital currency and monetary policy. Since 1996 there have been two main aspects regarding the impacts of digital currency on monetary policy. Tanaka (1996) stated that central banks will lose control over monetary aggregates and the foreign exchange rate, affecting money supply, and more seriously, potentially leading to a financial crisis. However, Ely (1996) argued
that the essence of digital currency is identical to that of fiat currency; therefore the monetary policy implications of digital currency are nil.

In order to emphasize the significance of digital currency, Berentsen (1997) pointed out the complementary role of digital currency for fiat currency. Additionally, Ali (2014) believed that threats to the stability of the current monetary system would be increased with the widespread use of digital currency in the future. Furthermore, Al-Laham et al. (2009) demonstrated the difficulties of digital currency in measuring the monetary aggregate, which would limit the ability of central banks to apply monetary policy via adjustments to money supply. Considering the underlying effects, the Bank of England (2014a) indicated that the potential risks of digital currency cannot be ignored in the long run. Specifically, conducting research into the Chinese market, Huang and Chen (2006) stated that digital currency can affect the money supply, and it will challenge the role of central bank of China in manipulating monetary policy. Meanwhile, Lei (2013) agreed that digital currency could bring a competitive money supply market, which could pose a significant challenge to the fiat currency issued by the central bank of China.

In contrast, Freedman (2000) claimed that the role of the central bank cannot easily be replaced by digital currency, even though the influence of digital currency are increasing. This opinion has been supported by Fang. Fang (2014) illustrated that it is unlikely that digital currency can completely replace traditional fiat currency; however, adjustments to monetary policy should be taken into consideration.

1.2.2. Aims and Objectives
In order to contribute to the academic discussion outlined above, this dissertation will conduct an in-depth analysis of digital currency in order to develop a framework for studying the impacts of digital currency. This paper examines how the proliferation of digital currency would affect the Chinese monetary system, particularly its fiat currency and monetary policy. Section 2 will start from the perspective of money demand, discussing the properties of digital currency and presenting a comparison of digital currency and fiat currency. In Section 3, an econometric model will be introduced to examine the relationship between digital currency supply and the velocity of money, followed in Section 4 by an analysis of the results and recommendations for Chinese monetary policy before concluding with a summary of the dissertation.
2. Literature Review

2.1. Definitions of digital currency

Since the rise of the Internet revolution in the 1990s, the definition of digital currency has evolved over time. The first definition of digital currency was given by the Basel Committee on Banking Supervision (BCBS), which referred to a prepaid payment terminal to store monetary value, which could be transferred by electronic devices with Internet access (BCBS, 1998). Subsequently, the European Central Bank and the Bank for International Settlement (BIS) published reports illustrating that digital currency is ‘an electronic store of monetary value on a technical device’, which can be used for making payments and investments without the engagement of a third party for transactions (European Central Bank, 2000; BIS, 2000). In 2008, the International Monetary Fund (IMF) reiterated the definition provided by the European Central Bank and proposed a simplified explanation of digital currency as an Internet-based medium of exchange (IMF, 2008).

2.1.1. Classifications of digital currency

Rooted in the various types of complex technology and methods, there are various classifications of digital currency in the financial market. For example, the IMF (2008) divided digital currency into two forms – hardware-based digital currency and software-based products depending on differences in existing terminals. Hardware-based digital currency is generally prepaid digital currency, represented as plastic cards issued by commercial banks, while software-based digital currency is primarily used in online remote payments. At the same time, X. Jiang (2010) specified digital currency into digital currency in cash within the form of commercial bank accounts. Furthermore, building on the IMF classification, Griffith (2014) further distinguished digital currency based on online and offline electronic payment systems.

This dissertation focuses on the impacts of digital currency on the Chinese monetary system; therefore, the classification of digital currency has been limited to two categories: those issued by financial institutions and those issued by non-financial institutions.

A typical example of digital currency issued by a financial institution is the smart cards which are predominantly used in face-to-face transactions. These include debit and credit cards and online banking within the platform of commercial banks. Without the need for explicit government controls, non-financial institutions can issue digital currency (IMF, 2008), providing the opportunity to create a new form of currency. Therefore, the second category, created by non-financial institutions on the
online platform, is an advanced form of digital currency that has excluded the role of commercial banks as a third party, with the aim of achieving peer-to-peer online transactions (Ali, 2014; Fung et al., 2014).

More specifically, the second category can be further divided into two parts: first, with the value of fiat currency from the online platform, but independent of the commercial bank; the second one is with the value of the new currency itself. The typical example here is Bitcoin, a new form of currency that cannot be measured in fiat currency but that does have an equivalent monetary value as a currency.

2.1.2. Incentives for creating digital currency

There is no doubt that one of the characteristics of digital currency is that it is tradable on a worldwide basis with no complex international boundaries. As a choice for e-commerce, the online digital currency platform presents a convenient method to make a payment directly without further confirmation or recognition (Al-Laham, 2009). Normally, traditional trading relies on a financial institution that can serve as a trusted third party. However, in the traditional currency system, trading volumes are limited by a minimum transaction size. In contrast, because digital currency is based on electronic payment processing on Internet platforms, there are no restrictions on small-sized transactions, which could reduce the payment costs (HM Treasury, 2015).

Therefore, the main initial incentive for private institutions or non-government organisations to issue digital currency is to reduce the costs of transaction. Especially for small business and individual investors, without the involvement of a third party, digital currency can be more efficient in dealing with payments by holding money online, which could result in lower transaction costs and information-gathering costs (Kumar, 2012; Bank of England, 2014a).

In the development of digital currency, another important characteristic is the learning spillover (Berentsen, 1997). This indicates that the technical growth has more influence on digital currency than on fiat currency. In this case, software-based digital currency can benefit individuals by improving their skills and knowledge of using the Internet and result in optimization of the use of money. With the expansion of Internet users, the value of digital currency will correspondingly increase (European Central Bank, 2012). Therefore, while the learning spillover effect is to generate increasing recognition of e-commerce and digital currency, it is likely to reduce the use of cash and the replacement effects to fiat currency will gradually become more apparent.
2.2. Digital currency and fiat currency

Fiat currency has played the crucial role of a medium for adjusting monetary policy on the monetary supply side. Therefore, in order to identify the impacts of digital currency, the first step is to compare the functions of digital currency and fiat currency.

2.2.1. Currency functions of digital currency

The globally accepted main functions of traditional currency include: medium of exchange, unit of account and store of value. To illustrate the influence of digital currency, it is important to discuss whether digital currency has the same functions as fiat currency.

2.2.1.1. Medium of exchange

From the perspective of a medium of exchange, there is no doubt that digital currency can be freely exchanged as a method of payment (Ali et al., 2014). Riza (2010) emphasised that digital currency can not only be exchanged in transactions, but can also be redeemed for cash, which represents the role of medium of exchange. Currently, there are various retail sectors, especially online retailing, willing to accept digital currency (Grinberg, 2012). However, not all commercial sectors accept digital currencies; a long time is still needed for it to be accepted in all markets compared with the globally recognised fiat currencies (Plassaras, 2013).

2.2.1.2. Unit of account

There is little evidence to prove the function of digital currency as a unit of account. Generally, according to the characteristics of digital currency, the majority of digital currencies still rely on fiat currency to reflect their market value. However, one of the special products is Bitcoin, issued by non-financial institutions and which is completely independent from any fiat currency. A unit of Bitcoin is counted as Bitcoin itself rather than as a unit of fiat currency (Nakamoto, 2008). It has already established a relatively complete unit which can be easily transferred into fiat currency. However, Goldeman (2014) claimed that without an intrinsic value, the high price of Bitcoin is a bubble. This view illustrates that the risk of this type of independent digital currency is comparatively high. Therefore, whether digital currency can be regarded as a unit of account in the same way as fiat currency can remains uncertain and requires further research in the future.

2.2.1.3. Store of value

Store of value refers to the fact that a given asset has value and ‘can be stored and retrieved in the future’ (Valdes-Benavides and Hernandez-Verme, 2014). Meanwhile, the increasing number of users
of digital currency show that it can satisfy people’s liquidity preference (Lee and Turban, 2001). With the demand for digital currency identified, the supply of digital currency will remain (Bank of England, 2014a). However, without its official admission, the value of digital currency would simply depend on the market preference (Bank of England, 2014b). Therefore, the demand for digital currency is uncertain.

Although the three functions of digital currency are not explicitly comparable with those of fiat currency, it cannot be denied that there are similarities between digital currency and fiat currency.

2.2.2. Trust and risk of issuer

Accompanying the similar functions in the properties of digital and fiat currency, the replacement effects of digital currency on fiat currency can also be taken into account.

Griffith (2014) demonstrated that the issuer of digital currency determines the effects of monetary policy. Meanwhile, HM Treasury (2015) and the Financial Action Task Force (2014) focused on the authority of the issuer, suggesting that the essential difference between digital currency and fiat currency is the involvement of a third party associated with central banks, and where control of the monetary policy conducted by the central banks lies. Specifically, fiat currency is based on the public assessment of the creditworthiness of the different currency issuers (European Central Bank, 2012), which means that the value of currency may vary according to their credit ratings. Currently, trust in the market is based on government legalisation, and the central bank plays a leading role in the economy as a trusted medium (Kiefer, 2000; Kumar, 2012). Meanwhile, in the traditional banking system, commercial banks are controlled by central banks to manage their capital operations and deliver efficient data transformation (Wagenhofer, 2003). However, without commercial banks involved as a transfer medium, direct transactions between two parties could generate lower costs. At the same time, compared with digital currency, it is difficult for commercial banks to avoid information asymmetries and market conflicts (Kiefer, 2000). Therefore, the complexity of the existing financial market would generate conflicts of trust, which in turn gives rise to the requirements of digital currency.

In addition, Riza (2010) suggested that digital currency is safer when compared to fiat currency. For example, lower-rated bank loans can result in default. Therefore, customers would consider online financial products more due to the security they offer (Lee and Turban, 2001). Taking risk and security into consideration, Nakamoto (2009) systematically proposed a system for electronic transactions without relying on a third party associated with trust, which is likely to diminish the conflicts of security.
and risk concerns. Compared with fiat currency issued by a trusted third party, the digital currency platform is not based on trust, but on cryptographic proof, known as a block chain network (Nakamoto, 2009). Therefore, this digital currency platform could reduce the default risk of the third party, fulfil a direct transaction between two parties, reduce the costs from a third party and lower the risk of trust.

2.3. Digital currency and money supply

The digital currency issued by financial institutions and non-financial institutions would have different impacts on fiat currency and monetary policy.

As an outcome generated by the financial institution, digital currency based on the commercial banks’ electronic platform would gradually reduce the free cash in the market (Liu and Xu, 2004). On the other hand, digital currency issued by non-financial institutions, which is independent of central bank controls and cannot be created by commercial bank loans, could alter the stock of fiat currency and consequently influence the monetary policy (Al-Laham et al., 2009).

In a traditional money system, central banks normally control the economy through monetary policy to adjust the supply of fiat currency. At the same time, central banks stimulate money supply by setting interest rates at such a level as to encourage commercial banks to create bank loans to meet market requirements (Bank of England, 2014b). Nevertheless, with the growth of digital currency supply, central banks’ assets and liabilities will be reduced, which will result in management problems. Specifically, the money reserve that the central banks uses to conduct the monetary policy will be altered (Fung, 2014).

Sharing a similar function to that of fiat currency, Dorn (1997) emphasised that digital currency could weaken the functions of monetary policy by affecting the velocity of money. Based on the personal computer, digital currency is accessible wherever there is Internet access, meaning that the velocity of money may be increased. Along with a reduction in fiat currency, a potential threat may arise from the increasing stock of digital money without legal control. Hence, Griffith (2014) and HM Treasury (2015) stated that the substantial effect of digital currency advancement would cause the elimination of monetary policy.

As previously mentioned, the properties of digital currency provide the possibility to replace fiat currency with increasing public incentives. Berentsen (1997) illustrated that digital money is expected to replace fiat currency, and may further affect the money supply. However, Ely (1996) argued that the
effects of digital currency are not significant to fiat currency; as a method of money circulation, there are no differences between digital currency and fiat currency, meaning that the monetary policy implications of digital currency can be ignored.

In order to identify whether the monetary system could be affected by digital currency, prior research has contributed by using different approaches in analysing its potential impacts. Pang (2003) explained money demand and money supply to show the effects of digital currency. Subsequently, Tang (2004) presented a mathematical model to support the relationship between money supply and money demand and further focused on the money multiplier in explaining the impacts of digital currency. With respect to digital currency, Zhou (2006) contributed to the velocity of money to illustrate the volatility of digital currency supply. In addition, Amromin and Chakravorti (2008) adopted an empirical analysis in studying the changes in demand for fiat currency within 13 countries, and concluded that the demand for fiat currency would fluctuate according to the prevailing interest rate rather than by the increasing availability of digital currency. Furthermore, X. Jiang (2010) demonstrated the disparity between digital currency and fiat currency in the supply and implementation of monetary policy.

To identify how the aggregate digital currency could affect the monetary system, the following section will employ econometric analysis in an attempt to explain the impacts of digital currency from money supply and velocity of money, respectively.

3. Methodology

3.1. Model assumption

Generally, central banks conduct monetary policy by adjusting money supply. Therefore, to understand the impacts of digital currency it is important to examine the changes in money supply. However, the measurement of money supply varies between countries and does not have a standard measurement. In order to illustrate the effects of digital currency in the Chinese market, a Chinese money measurement has been applied in this model. The Chinese government uses $M_0$, $M_1$ and $M_2$ to evaluate the money supply. Their definition has been provided by the People’s Bank of China (2013):

$M_0$: notes and coins in circulation outside of the banking system, including cash held by institutions and individuals.

$M_1$: narrow money supply, includes $M_0$ plus the current deposits of enterprises and public institutions.
M₂: broad money supply, includes M₁ plus the fixed deposits of enterprises and public institutions and household saving deposits.

Due to the characteristics of digital currency, the Chinese government has not yet proposed a specific criterion to distinguish digital currency from fiat currency. Hence, it is difficult to directly compare the amount of fiat money and digital currency. In calculating the digital currency supply, insufficient data might contribute to problems of double counting and lead to inaccurate statements of the supply of fiat currency and digital currency. To avoid inaccuracies in measuring the supply of digital currency, Zhou (2006) suggested applying velocity of money to analyse its impacts on China’s monetary system. For the same purpose of avoiding inaccurate statement of digital currency measurement, this dissertation focuses on the velocity of money.

It is necessary to clarify the relationship between money supply and velocity of money. The quantity theory of money offers a coherent framework to analyse the effects, which illustrates a relationship between the supply of money and nominal output. Fisher and Brown (1911) developed the equation of exchange:

\[ M_t V_t = PY \]

The equation states that the total amount of money in circulation (Mₜ) multiplied by the number of times that that money is circulated (the velocity of money, Vₜ) is equal to the level of nominal expenditures (PY), indicating an inverse relationship between the money supply and the velocity of money for a given nominal output. In other words, an increase in money supply could be offset by a decrease in the velocity of money, resulting in an unaltered outcome in the nominal expenditure. This model considers PY as the nominal output that could be counted as the Gross Domestic Product (GDP). According to the equation, if the aggregate money supply remains constant, the velocity of money is predicted to bring indirect effects to the money supply. Therefore, the velocity of money is calculated by GDP/Mₜ.

3.2. Methodology, data sources and data description

3.2.1. Model selection

In previous research, Liu and Xu (2004) and Zhou (2006) adopted quantitative models in analysing the monetary velocity and digital money, discussing whether digital currency could effect changes to the velocity of money. Jiang (2012) explored the indicators based on the assumption of Zhou (2006) and proposed correlations of quantifying electronic currency level. To adjust the current state of China’s
economy, this model would combine modifications with variable selections and date collections. In this chapter, cointegration test will be applied in the long-term equilibrium analysis; meanwhile, the unit root test and the error correction model will be adopted in the short-term non-equilibrium analysis.

3.2.2. Data collection

Chinese GDP data are the accumulated sum of quarterly and monthly rates. Collecting quarterly or monthly data directly would result in a multi-correlation problem. In order to guarantee the consistency of data, annual data are recommended in the model analysis. Meanwhile, Chinese market data have been systematically recorded since 1978. Therefore, the sample data have been selected from 1978 to 2013.

According to the three money classifications $M_0$, $M_1$ and $M_2$, the velocities of money in China are correspondingly defined as $V_0$, $V_1$ and $V_2$, which can be calculated by $\text{GDP}/M_t$. In this model, the dependent variables $V_0$, $V_1$ and $V_2$ will be discussed separately with the selected independent variables.

3.2.3. Operationalisation of independent variables

In order to quantify how changes in the aggregate digital currency could influence the amount of fiat currency, four independent variables have been selected from previous research: (1) cash ratio, (2) electronic currency level, (3) financial electronic level and (4) interest rate.

(1) Cash ratio (CR) = $M_0/M_2$

Along with the development of digital currency, the effects of digital currency on fiat currency can be reflected in different ways. The most significant impact on the financial market would be a reduction in the use of cash, which is defined as the replacement effect. While the digital trading system brings efficiency for market participants to exchange and invest, the convenient access to cash and the increasing expectation of withdrawing cash would reduce the preference of the public to hold money (Qu, 2013). On the other hand, while a decrease in the use of physical cash would result in a decrease in the velocity of money, the complementary effects of digital currency could compensate for this fall in the velocity of money.

The definition of cash ratio is presented by Zhou (2006), with a different explanation here to that of the financial measurement of company liquidity. As the Chinese market defined, $M_1$ is used to measure the broad currency and $M_0$ contains cash circulating in the market. Therefore, the percentage of cash
in broad money \((M_0/M_2)\) would be an indicator to evaluate the relationship between fiat currency and digital currency. In summary, due to the replacement effect of digital currency for cash, an expansion of digital currency would bring about a reduction in cash, corresponding in practice with a lower velocity of money. On the other hand, the complementary effect of cash could diminish the replacement effects to some extent. Especially, \(V_0\) is used to measure the velocity of cash in circulation; therefore, the complementary effect is more noticeable in \(V_0\).

(2) **Electronic currency level (EC)** = \((M_1-M_0)/M_2\)

Zhou (2006) introduced a concept of money supply liquidity to quantify the impacts of digital currency, which is defined as \(M_2/M_1\). However, Z. Jiang (2012) adopted grey relational analysis and concluded that as the development of digital currency in China remains in its initial stages; digital currency would have more impacts in bank current deposits, which would be more accurately measured by \(M_1-M_0\). According to this, the model replaces the numerator \(M_1\) with \(M_1-M_0\) as an alternative. Dividing \(M_1-M_0\) by aggregate broad money \((M_2)\) provides the electronic currency level as the second indicator.

In summary, the digital currency issued by financial institutions would enhance the utilisation of bank current deposits. By facilitating the usage of bank current deposits, digital currency would cause fluctuations in the velocity of money.

(3) **Financial electronic level (FE)** = \((M_2-M_0)/M_2\)

There is an indirect impact on money demand and the financial electronic level. A more complete and regulated financial system brings a higher velocity in the market. Diversifying asset allocation online by using digital currency, people are less likely to hold money in their bank accounts and turn to the higher interest gains from Internet investment rather than the lower interest rates offered by the commercial banks. As a consequence, with the accelerated financial electronic level, the velocity of money will be constantly increased.

The equation for the financial electronic level was proposed by Ai and Fan (2002). Ai and Fan (2002) illustrated that the higher the weight of non-cash in the financial system, the higher the degree of financial electronic level would be. Therefore, a decrease in the free flow of cash would result in an increase of \(M_2-M_0\), and generate a higher degree of financial electronic level. According to the practical statistics, the degree of financial electronic level has been improved since the 1990s, along with a decline in the velocity of money (Zhou, 2006).
Since the monetary side of the economy was more volatile than the real side in the late 1970s, Poole (1970) suggested that interest rate policies are more effective at stabilising economies. In recent years, central banks have favoured interest rate rules rather than money supply policies to conduct monetary policy. As a consequence of monetary policy, a lower interest rate would reduce the possibility of holding money rather than devoting it to investment and consumption, so that the demand for money would be increased and would eventually affect the money supply and velocity of money. Tang (2005) believed that the interest rate could be a better measure for testing monetary policy, because the use of digital currency introduces a lot of difficulties to the measurement of currency, which would diminish the effects of adjusting the money supply. Meanwhile, Liao and Tapsoba (2014) claimed that interest rate liberalisation slowed down the velocity of money. Therefore, in this model, the interest rate has been introduced as a price indicator of money that could have impacts on the velocity of money.

3.3. Testing China’s velocity of money model

This dissertation summarises the relationship between independent variables and the velocity of money as follows:

\[ V_t = c + \alpha CR + \beta EC + \gamma FE + \delta I \]

Notes: \( V_t \) is velocity of money in different time period; \( CR \) is cash ratio; \( EC \) is electronic currency level, \( FE \) is financial electronic level; \( I \) is interest rate; \( c, \alpha, \beta, \gamma \) and \( \delta \) are coefficients.

The right side of the equation shows the indicators reflecting the digital currency supply to the fiat money supply. The left side of the equation is the velocity of money. Therefore, the equation links the digital currency supply and velocity of money together.

The following section uses the statistical software STATA and EViews to evaluate the relationship based on the application of Chinese data.

3.3.1. Unit Root Test

In order to ensure a valid assumption of the model, it is important to include stationary data. Therefore, the Unit Root Test is firstly introduced to classify the stationary data (Bo, 2008). This model applies the augmented Dickey-Fuller test to test a unit root, which aims at testing whether the time
series variables are non-stationary or not. Taking the logarithm of velocity of money, cash ratio, electronic currency level, financial electronic level and interest rate respectively, the testing results are listed as follows:

Table 1: Results of Augmented Dickey-Fuller Test

<table>
<thead>
<tr>
<th></th>
<th>V0</th>
<th>V1</th>
<th>V2</th>
<th>CR</th>
<th>EC</th>
<th>FE</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value</td>
<td>0.0039**</td>
<td>0.0002**</td>
<td>0.0001**</td>
<td>0.0065**</td>
<td>0.0003**</td>
<td>0.0077**</td>
<td>0.0004***</td>
</tr>
<tr>
<td>Result</td>
<td>Stationary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *** significant under 1% level.

By adjusting the setting of intercept and trend, the result show that the variabels are stationary in the 1st difference.

3.3.2. Cointegration Test

(1) Ordinary Least Squares (OLS) Method

In the cointegration test, the logarithmic form of velocity of money, cash ratio, electronic currency level, financial electronic level and interest rate are further applied in the regression test by using the OLS method. In order to discuss the long-term effects, each variable is associated with the time trends, so the intercept is included. The model can be defined as follows:

\[
LV_0 = 4.2360 + 0.0642LCR + 0.9313LEC + 4.7072LFE + 0.1078LI
\]

\[
(4.0391) \quad (0.2341) \quad (4.8725) \quad (1.9719) \quad (1.7069)
\]

\[
R^2 = 0.7583
\]

\[
LV_1 = 3.4708 + 0.7847LCR + 0.1912LEC + 4.3325LFE + 0.0788LI
\]

\[
(3.3637) \quad (2.9076) \quad (1.8447) \quad (1.7682) \quad (1.2682)
\]

\[
R^2 = 0.6876
\]

\[
LV_2 = 3.8267 + 0.9742LCR + 0.9175LEC + 3.8112LFE + 0.0818LI
\]

\[
(3.6028) \quad (3.5066) \quad (8.2417) \quad (1.5765) \quad (1.2792)
\]

\[
R^2 = 0.9470
\]

Notes: () are standard errors.
(2) Stationarity Test of Residual

Table 2: Results of Stationarity Test of Residual

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Independent Variable</th>
<th>P-value</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0</td>
<td>Residual</td>
<td>0.0004***</td>
<td>Stationary</td>
</tr>
<tr>
<td>V1</td>
<td>Residual</td>
<td>0.0005***</td>
<td>Stationary</td>
</tr>
<tr>
<td>V2</td>
<td>Residual</td>
<td>0.0005***</td>
<td>Stationary</td>
</tr>
</tbody>
</table>

Note: *** significant under 1% level.

The stationary result of the residual test provides the cointegration relationship between \( V_t \) and independent variables. In other words, the residuals do not cause a problem and there is a long-term equilibrium relationship.

3.3.3. Error Correction Model

According to Engle and Granger (1987), if there is a cointegration relationship, the Error Correction Model can be adopted in explaining the non-equilibrium relationship in the short term. It has been proved that the existing cointegration relationship between variables in the OLS model and the residuals are stationary; therefore, the Engle-Granger approach can be adopted to establish the Error Correction Model to test the short-term effects.

To establish the Error Correction Model, firstly, differentiate \( LV_0, LV_1, LV_2, LCE, LFE, LI \), and use the Engle-Granger approach to regress with residual \( E(-1) \), followed by adoption of stepwise regression to simplify the model and eliminate the variables with an insignificant P-value in the OSL Model.

The Error Correction Regression functions can be specified as follows:

\[
\Delta LV_0 = -0.0237 + 0.3248\Delta LEC + 3.7229\Delta LFE + 0.0326\Delta LI - 0.2995E(-1)
\]

\( R^2 = 0.5179 \)

\[
\Delta LV_1 = -0.0396 - 0.1128\Delta CR - 0.4536\Delta LEC - 0.1822\Delta LFE - 0.2762E(-1)
\]

\( R^2 = 0.5179 \)
\[ R^2 = 0.4049 \]

\[ \Delta V_2 = -0.0392 + 0.1238 \Delta CR + 0.3180 \Delta EC - 0.2771 \Delta E(-1) \]

\[ R^2 = 0.4730 \]

Notes: () are standard errors.

Table 3: Summary of Coefficients of Model Results

<table>
<thead>
<tr>
<th></th>
<th>Cash Ratio (CR)</th>
<th>Electronic Currency Level (EC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>Engle-Granger</td>
</tr>
<tr>
<td>V0</td>
<td>0.0642</td>
<td>-</td>
</tr>
<tr>
<td>V1</td>
<td>0.7847***</td>
<td>0.1912*</td>
</tr>
<tr>
<td>V2</td>
<td>0.9742***</td>
<td>0.9175***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Engle-Granger</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0</td>
<td>0.9313***</td>
<td>0.3248**</td>
</tr>
<tr>
<td>V1</td>
<td>0.1912*</td>
<td>-0.4536***</td>
</tr>
<tr>
<td>V2</td>
<td>0.9175***</td>
<td>0.318***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Financial Electronic Level (FE)</th>
<th>Interest (I)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>Engle-Granger</td>
</tr>
<tr>
<td>V0</td>
<td>4.7072*</td>
<td>3.7229***</td>
</tr>
<tr>
<td>V1</td>
<td>4.3325*</td>
<td>-1.1822</td>
</tr>
<tr>
<td>V2</td>
<td>3.8112</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: *significant under 10%, **significant under 5% level, ***significant under 1% level.

3.4. Empirical Analysis

3.4.1. Cash ratio

Overall, cash ratio is positively correlated with the velocity of money in the long term. In the OLS model, the coefficient in cash ratio for each level of velocity of money is less than 1, which implies that there is no accelerating effect of cash ratio on the velocity of money. Namely, a decrease in cash ratio would be accompanied by a relatively lower velocity.

As a complement of cash, digital currency could compensate the loss of velocity of money caused by the reduced use of cash. The coefficient in V0 close to zero indicates that the possible decrease in cash
usage could have insignificant impacts on $V_0$, which would be offset by the complementary digital currency. In contrast, the replacement effects of $V_1$ and $V_2$ override the complementary effect, which is shown as a coefficient with a value of around 1. However, with the development of digital currency in China remaining at an early stage, the replacement effects of cash are not yet significant. With increasing popularity of digital currency as an alternative to cash, acceleration to the velocity of money would be increased.

With an insignificant P-value, $V_0$ has been continuously discarded in the Error Correlation Model, which indicates that the complementary effect of digital currency in $V_0$ is more significant than the replacement effect. In a short-term analysis, the Error Correlation Model provides a negative coefficient of cash ratio in $V_1$, and $V_2$ is closer to 0. Namely, the compensation effect of digital currency of $M_1$ and $M_2$ are significant in the short term. Especially for $V_1$, a decrease in cash would result in an increase in the velocity of money.

In the long term, the cash in circulation would not simply be replaced by digital currency so that the cash ratio would be reduced by an amount corresponding to the acceleration in the rate of velocity of money.

3.4.2. Electronic currency level

The OLS model shows a substantial result of the electronic currency level in the long term, especially in $V_0$ and $V_2$, which present coefficients of nearly 1 as well as significant P-values. On the other hand, the Engle-Granger model shows significant P-values, but relatively lower coefficients in the short term. The opposite result in $V_1$ indicates that digital currency would have negative impacts on the velocity of $M_1$ in the short term. This could be caused by a public liquidity preference for holding digital currency rather than fiat money (Zhou, 2006). In the short run, holding digital currency is more convenient than the fiat currency and further reduces transaction costs; therefore, people would be more willing to hold digital currency. In addition, this preference can increase the time spent holding money, thereby reducing the velocity of money.

Furthermore, the majority of digital currency exists in bank current deposits, which have been used as the daily tool for transaction and payment. Moreover, fluctuations in bank current deposits are highly common in practice. Meanwhile, as a basic part of $M_1$, the velocity of bank current account varied the velocity of $M_1$ as well. Therefore, the impact of digital currency reflected on $V_1$ is negative.
Overall, the model results for electronic currency level suggest that the digital currency could have considerable effects on the velocity of money in the long term compared to with the short term.

3.4.3. Financial electronic level
The model results display a positive correlation between velocity of money and the financial electronic level. In the long term, the financial electronic level makes the greatest contribution in accelerating the velocity of money, which can notably be found in the relevant coefficients around 4. In comparison, the impact of the financial electronic level is substantial in \( V_0 \), which is associated with the advancement of digital currency. Along with a higher financial electronic level, cash in circulation would be accelerated in different forms of digital currency so as to accelerate \( V_0 \). Furthermore, the coefficient for each velocity of money is decreasing from \( V_0 \) to \( V_2 \). This situation could be a reflection of trust in the issuer. The issuer of digital currency has not yet gained sufficient public recognition, so that the influence in velocity of \( M_2 \) is not as significant as the velocity of \( M_0 \).

The short-term effects can only be revealed from \( V_0 \) in the Engle-Granger Model, which presents a high coefficient of 3.44. This could be caused by the substantial exchange medium function of digital currency instead of the function of storage of value. As an advanced payment system that has enhanced the development of currency exchange currency across time zones and geographical borders, people engage in digital currency as an exchange medium in real-time payments and transactions, which would be more apparent in the short term.

Generally, the relatively high coefficient in the financial electronic level indicates that the degree of financial electronic level is considerable. In recent years, the financial electronic level has been stimulated by financial innovation. Meanwhile, the increasing rate of the financial electronic level is accelerating year by year and is predicted to expand further in the future.

3.4.4. Interest rate
From the model results, the interest rate is positively correlated with the velocity of money. Although there is inefficient theory to support the positive relationship, it has been proved that there is a negative relationship between money supply and interest rate (Keynes, 1965). Referring to the equation of exchange, the inverse relationship between money supply and velocity of money could be a possible explanation for the outcome of the positive relationship; namely, an increase in interest rate would result in a higher velocity of money.
Additionally, the relatively low coefficient of interest rate could suggest that the effect of interest rate on the velocity of money is not significant. The low coefficient implies that the development of e-commerce improves the process of interest rate liberalisation, as well as diminishing the impacts on velocity. Furthermore, the interest rate has been abandoned in the Error Correlation Model, due to the insignificant P-value of interest rate in the OLS model.

The result is also consistent with Amromin and Chakravorti’s (2008) finding that there are no significant effects of changes in interest rate progress resulting from financial innovation that would affect the fiat currency.

In the China market, the interest rate is inelastic and is controlled by the Chinese central bank so that it cannot truly reflect the functions of interest rates (Liu and Xu, 2003). Hence, the long-term effect of interest rate on velocity is insignificant.

3.5. Limitations

Firstly, the development of digital currency is still in its early stage; hence, the data are limited to the period from 1978 onwards, 1978 being the year of the reform and opening-up policy in China and signalling the start of financial innovation. Meanwhile, the monthly and quarterly data are unavailable; therefore, the number of samples in each variable is limited to 35. Because the samples are insufficient, possible outliers would be generated from the undefined outside factors that could result in biased outcomes.

Secondly, from the perspective of quantification in the independent variables, the measurement of aggregate digital currency has not been discovered in a recognised method. Therefore, indirect indicators have been introduced in this model. At the same time, each indicator may not be explicitly defined. In this case, potential multicollinearity problems may not be covered in this model.

Thirdly, for the reason that there are limited researches focused on quantifying the aggregate digital currency, this model is based on the previous studies led by Zhou (2006), which adopted common regression models including the OLS model and continuously applied the Error Correlation Model. However, these models are not specially aimed at studying the relationship between those variables, so they would not guarantee precise results.
3.6. Model Summary

This model analysis has discussed the indirect factors reflecting the supply of digital currency that could have impacts on China’s velocity of money. The four indicators used to quantify the effects of digital currency on the changing amount of fiat currency were: cash ratio, electronic currency level, financial electronic level and interest rate.

In general, the electronic currency level and the financial electronic level have significant impacts on the velocity of money. Nevertheless, the interest rate does not play an important role in altering the velocity of money, and the opposite results given by the two models have not provided sufficient information to illustrate their relationship with the velocity of money.

In terms of digital currency, the substitution effect of digital currency on free cash and technological growth in the financial sector explained a view that the enhanced impacts of digital currency would generally affect the money supply indirectly by influencing the velocity of money. While there is a relationship between the supply of digital currency and the velocity of money, China’s monetary system would be challenged.

4. Results and Policy Recommendations

Section 3 proved that digital currency is gradually changing the structure of the money supply and the velocity of money in China in either the long term or short term. In other words, the extensive use of digital currency will bring considerable challenges to fiat currency, which could further affect the traditional Chinese monetary system and monetary policy.

Firstly, the classifications of digital currency cause inaccurate measurement of monetary aggregate, bringing uncertainties and limitations to the study of the relationship between aggregate digital currency and money supply in the model. Al-Laham et al. (2009) suggested that ‘the decreased ability to measure monetary aggregates will limit the central bank’s ability to conduct open market operations and target the money supply’. In this case, the increasing use of digital currency would bring difficulties to money measurement, which would create challenges in controlling money supply. Therefore, money measurement should be considered in discussing the monetary policy of China. By evaluating the independent variables in Section 3, the central bank of China needs to control the aggregate money supply indirectly. The model results suggest that the velocity of money will be affected by variations in
aggregate digital currency. Hence, it is important to clarify the form of digital currency and fiat currency from the standardised money measurement.

Secondly, it was discussed in Section 2 that the issuers of digital currency could be divided into either financial institutions or non-financial institutions; they are challenging the independent role of central bank to conduct monetary policy (Brentern, 1997; Jiang, Z., 2012). As Liu (2010) mentioned, digital currency is reforming the behaviour of the financial system and the public. On the money demand side, rapid economic development has encouraged development of the Chinese e-commerce market. Meanwhile, as the intermediaries of e-commerce, the incentives for using digital currency stimulate the constant growth of the digital currency market. Specifically, sharing similar functions with the fiat currency, the structure of money demand becomes complicated with the emergence of digital currency, which has the potential to create uncertainty so as to bring difficulties to policy transmission (Valdes-Benavides and Hernandez-Verme, 2014). For this reason, the authority regulations and supervisions are required to adjust the demand of digital currency when conducting monetary policy.

Furthermore, there are disparities in the long-term and short-term effects in the relationships of the model variables. Generally, the development of digital currency is still in its early stage; therefore, monetary policy would not be easily affected by digital currency in the short term. On the other hand, the development of digital currency mainly contributes to the long-term effects in money supply as well as the velocity of money, indicating that the central bank of China needs to conduct its monetary policy with one eye on the future.

In order to improve monetary policy, future research is required to further identify the replacement effects of the fiat currency and the eliminating role of fiat currency. It is suggested that the central bank of China combines quantity and price tools (PBOC, 2013). Meanwhile, promoting the security of Internet finance and the reliability of digital currency, Chinese authorities need to continue studying the characteristics and further identify the functions of digital currency. At the same time, it is necessary to improve pertinency in financial system regulations and digital market supervision, along with optimising the financial market, preventing systematic risks and standardising evaluation systems. Overall, the central bank of China needs to continuously make structural economic adjustments and upgrade its transformation to keep up with the progress of digital currency.
5. Conclusion

This dissertation set out to explore how digital currency may affect Chinese monetary system in discussing the properties of digital currency, including the definition of digital currency and a comparison with the fiat currency. Analysis of the impacts of digital currency was conducted by establishing an econometric model to clarify the money supply and velocity of money in the Chinese market specifically. It concluded with Chinese monetary policy recommendations.

As mentioned in Section 2, e-commerce has increased the popularity of digital currency. Along with the diversified forms of digital currency, its gradual advancement enables digital currency to share similar functions to those of fiat currency. At the same time, potential risks of trust have been taken into consideration. Consequently, the possible replacement effects of fiat currency with digital currency led the discussion on the supply of money and velocity of money.

Although it is too early to state whether digital currency will eventually replace fiat currency, there is no denying that the velocity of fiat currency will become more flexible whilst widespread digital currency will have a considerable impact in the Chinese market over time, according to the econometric model in Section 3. To avoid interpretation problems in the classification and measurement of digital currency, the different levels of Chinese monetary classification bring considerations of three classes of velocity of money, which has been regarded as a tool to clarify the relationship between digital and fiat money supply. In the model assumption, the aggregate digital currency is supposed to be reflected in the changes in cash ratio, electronic currency level, financial electronic level and interest rate, which will cause fluctuations in the velocity of money. Consequently, the model results demonstrate that the effects of digital currency would be different in the short- and long-terms, which would pose challenges to the implementation of monetary policy. However, the insufficient data and undefined classifications limited the results.

In conclusion, it cannot be denied that the threats of digital currency could cause an extensive and complex situation for the Chinese central bank in its conducting of monetary policy, especially when the volume of digital currency is underestimated by the central bank. With the proliferation of digital currency, the independent role of the Chinese central bank is predicted to come under pressure, which may further limit monetary policy transmission. Therefore, the relevant policy recommendations suggest that the central bank of China standardises its money measurement and improves authority regulations and controls so as to upgrade the financial system through understanding digital currency.
Bibliography


344
Available at: http://www.wsj.com/articles/the-revolutionary-power-of-digital-currency-1422035061

Available at: http://www.ce.cn/celt/wyry/201306/18/t20130618_24487125.shtml

Available at: http://tech.sina.com.cn/i/2013-04-15/14138240605.shtml

Available at: https://www.cnnic.cn/hlwfzyj/hlxwzb/hlwtjbg/201207/P020120723477451202474.pdf

Available at: http://www.ebrun.com/20150202/123162.shtml


Available at: https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf


Kumar, M. 2012. Relationship between types of trust and level of adoption of Internet banking. Problems and Perspectives in Management. 10(1). pp.82-92.


Qu, Y. 2013. The impacts of digital currency on interest rate. North Economy. 7(2). pp.3-4.
Available at: http://www.justice.gov.tr/e< journal/pdf/basic_pr%C4%B1nciples.pdf

[Accessed 8 March 2014].
Available at: http://www.fincen.gov/news_room/rp/files/e-cash.pdf


[Accessed 8 March 2014].

Available at: http://www.fatf-gafi.org/media/fatf/documents/reports/virtual-currency-key-definitions-and-potential-aml-cft-risks.pdf

