Erasmus
School of
Economics

Erasmus University Rotterdam

Erasmus School of Economics

MSc. Economics & Business

Master's Specialisation—Financial Economics

Master Thesis

Responsible Investing: Empirical Analysis into Exchange-traded Funds Performances and its Resilience over COVID-19 Exogenous Shocks

Name: Ong Hong Shen Jevon

Student ID number: 545714jo

Supervisor: Prof. Dr. Onno Steenbeek

Second assessor: Dr. Jan Lemmen

Date final version: 4 August 2022

ERASMUS UNIVERSITEIT ROTTERDAM

JEL Classification : C38, E44, G11, G12, G14, M14

Keywords: Exchange-traded funds; Multifactor Analysis; Performance Measures; SFDR; Event Studies

Erasmus University Rotterdam
Making Minds Matter

Acknowledgements

I want to thank the two pillars of my life: family and friends. Thank you, mummy and papa, for supporting me with my studies abroad. Five years in, I have become more aware of the personal cost of an education abroad. Not just financial costs, but the costs of my absence, that I put on my grandparents in the last years of their life. As I transition towards an international career, something I have yearned for all my life, I will bear this heavy burden of personal costs. With some luck, I hope to be able to return to Singapore in my later years to uplift our family and country to even greater heights.

Second, I would like to thank my supervisor, Onno Steenbeek, for his support and guidance throughout my master's thesis. His valuable feedback and suggestions navigated me in the right direction whenever necessary. In addition, his asset management seminar confirmed my interest in a career in investment.

Third, thank you, Christoph and Pierre. You have been instrumental in guiding me not just through my academics but also in my professional career. Let us keep punching above our weights, and, in time, we shall stand on the shoulders of legendary financiers.

Last, I hope this thesis and the results found are enlightening and useful for the readers in their investments.

Jevon H.S. Ong
August 2022

.

tl;dr The clean portfolio (SFDR articles 8 and 9 ETFs) outperformed the dirty portfolio (SFDR article 6 ETFs) and was more resilient during COVID-19 by being less negatively impacted in their returns. The study of this thesis focused on ETFs in Europe with Euro denomination.

Abstract

This master’s thesis adds to the literature on responsible investment performance differences of clean portfolios (SFDR article 8 and 9 sustainable ETFs) by assessing the risk-adjusted returns. It also adds to the literature on performance resilience during COVID-19 by examining the impact on returns and its factor sensitivities. Employing the multifactor models, the clean portfolio performed better than the European market and comparatively better than the dirty portfolio (SFDR article 6 non-sustainable ETFs). The use of performance measures further supports the findings. They scored better 75% of the time for a clean portfolio than a dirty portfolio. The former performed better using risk measures with the following components: full volatility, downside deviation and drawdowns. Clean portfolio not only has better performances during COVID-19 and in the medium-term, but it also has outperformance over the long run using the macroeconomic factor model. Regarding resilience, multifactor models showed that the clean portfolio of ETFs was comparatively less impacted than the dirty portfolio during COVID-19. The answer was inconclusive on market factor when comparing the estimates of a clean portfolio against a dirty portfolio. In size factor, the factor estimates gave inconclusive answers as well. A meaningful interpretation of these factor loadings can still be drawn from the difference as a regressand. Both size and value factors were negative and significant in the coefficient estimate. The results show that a clean portfolio is more exposed to small-cap and growth stocks. Lastly, adding two quality factors from Fama-French and Carhart’s momentum factor had no significance.

Table of Contents

1. Introduction 1

1.1 General Overview of Sustainable Investing 2

1.2 Sustainable Financial Disclosure Regulation (SFDR) 4

1.2.1 SFDR Article 9 funds (sustainable investment products) 4

1.2.2 SFDR Article 8 funds (ESG-integrated products) 4

1.2.3 SFDR Article 6 funds (all other funds; non-sustainable products) 5

1.3 Area of research, its originality, the relevance, and its objective 5

2. Literature review 7

2.1 Performance difference of investments with higher ESG 7

2.2 Resiliency of article 8 funds throughout COVID-19 pandemic as an exogenous shock 10

3. Conceptual Framework 13

4. Data 17

4.1 Data Collection 17

4.1.1 Dependent Variable 18

4.1.2 Independent Variables (Part 1 CAPM, Multifactor Analysis) 18

4.1.3 Independent Variables (Part 3 Expanded Multifactor analysis) 18

4.1.4 Independent Variables (Part 4 Macroeconomic Factor Analysis) 19

4.2 Dataset Preliminary Checks (Postestimation Statistics) 20

4.2.1 Assumption 1: Errors have zero mean [$E(u_t) = 0$] 20

4.2.2 Assumption 2: Homoskedasticity [$Var(u_t) = \sigma^2 < \infty$] 21

4.2.3 Assumption 3: No autocorrelation [$Cov(u_i.u_j) = 0$, for $i \neq j$] 21

4.2.4 Assumption 4: Residuals are normally distributed [$u_t \sim N(0, \sigma^2)$] 22

4.3 Factor Loadings' Summary Statistics (Multicollinearity Checks) 23

4.3.1 Part 1's CAPM and multifactor analysis factor loadings (monthly returns) 23

4.3.2 Part 3's Expanded multifactor analysis factor loadings (daily returns) 24

4.3.3 Part 4's Macroeconomic factor analysis factor loadings (quarterly returns) 25

5. Methodology 26

5.1 Responsible Investing—Portfolio Performance Differences 26

5.1.1 CAPM and Multifactor Analysis (Monthly Returns) 26

5.1.2 Performance Measures (Monthly Returns) 27

5.2 Responsible Investing—Resilience over COVID-19 (Event Study) 31

5.2.1 Single Factor and Multifactor Analysis (Daily Returns) 31

5.2.2 Macroeconomic Factor Analysis (Quarterly Returns) 32

6. Results 33

6.1 Part 1 CAPM, Multifactor Analysis (Monthly Returns) 33

6.1.1 Summary statistics 33

6.1.2 Regression results 34

6.2 Part 2 Performance Measures (Monthly Returns) 36

6.3 Part 3 Expanded Multifactor Analysis (Daily Returns) 39

6.3.1 Summary statistics 39

6.3.2 Regression results 41

6.4 Part 4 Macroeconomic Factor Analysis (Quarterly Returns) 44

6.4.1 Summary statistics 45

6.4.2 Regression results 46

7. Discussion 48

8. Conclusion 52

8.1 Limitations of Research 53

8.2 Recommendations for Further Research 54

8.3 Concluding Remarks 55

9. References 56

10. Appendices 62

Appendix A. Supportive literature of sustainability performance difference and resilience 62

Appendix B. Opposing literature of sustainability performance difference and resilience 65

List of Tables

Table 1. ETF summary (number of ETFs per frequency) ………………………………………… 17

Table 2. Postestimation statistics………………………………………………………………….. 20

Table 3. Summary statistics and checks of multicollinearity (part 1 analysis) …………………… 24

Table 4. Summary statistics and checks of multicollinearity (part 3 analysis) …………………… 24

Table 5. Summary statistics and checks of multicollinearity (part 4 analysis) …………………… 25

Table 6. Summary statistics of monthly returns, with percentages as unit of measure …………... 33

Table 7. Regression results of monthly returns, with percentages as unit of measure …………… 35

Table 8. Performance measures based on monthly returns between January 2017 to May 2022 … 36

Table 9. Summary statistics of daily returns, with percentages as unit of measure ……………… 39

Table 10. Regression results of daily returns, with percentages as unit of measure ……………… 43

Table 11. Summary statistics of quarterly returns, with percentages as unit of measure ………… 45

Table 12. Regression results of quarterly returns, with percentages as unit of measure ………… 46

Table 13. Overview of literature supportive of responsible investing …………………………… 62

Table 14. Overview of literature opposing responsible investing ………………………………... 65

List of Figures

Figure 1. Types of Sustainable Investing Strategies ………………………………………………… 3

Figure 2. Conceptual Overview of Analysis ……………………………………………………… 13

Figure 3. Conceptual Framework of Part 1 Multifactor Analysis (Monthly Returns; Indices as Factor Proxies) ………………………………………………………………………………. 14

Figure 4. Conceptual Framework of Part 3 Expanded Multifactor Analysis (Daily Returns) ……. 15

Figure 5. Conceptual Framework of Part 4 Macroeconomic Factor Analysis (Quarterly Returns). 16

Figure 6. Monthly Returns and Monthly Sharpe Ratio Movements ……………………………… 34

Figure 7. Monthly Drawdown Movements ……………………………………………………….. 38

Figure 8. Daily Returns Movements, in percentages ……………………………………………... 41

Figure 9. Quarterly Returns Movements, in percentages ………………………………………… 45

1. Introduction

Sustainable investing, an umbrella term for investments that are compatible with sustainable development, was once viewed as one of the differentiating factors an investment manager can have against the others. In recent years, that is no longer the case. Increasingly, many investors are concerned about the environmental, social and governance (ESG) challenges faced by this generation and future generations. Sustainable investing is now viewed not as a differentiating factor but as an essential factor to ensure investment activities do not add to the ESG challenges faced.

Investment activities require excess returns above the market, but do these sustainable investment activities eat into the investors' returns? A wealth of academic research provided an inconclusive answer. Proponents of such sustainable investment activities point to empirical evidence of sustainable investment activities providing excess returns above the market. Others point to no significance of underperformances against the market, a fact that no excess financial returns are acceptable because investors derive benefits by enabling sustainability solutions instead. Opponents of such sustainable investment activities point towards evidence of outperformance for portfolios that do not limit their investment universe by sustainability. Those excess returns, they argued, can then be channelled by the individual or institutional investors towards sustainable activities, such as investing in green projects. A case in point from these opponents stems from the highly cited research paper on sin stocks' outperformance (Hong & Kacperczyk, 2009).

Furthermore, the question of divestment or continued stewardship in companies with low sustainability or low social responsibility has been subjected to much debate. On the one hand, proponents argued for divestment of such firms due to inconsistent alignment with investors' values or investment objectives. On the other hand, opponents argued for continual investment in such companies to enact change through stewardship, such as proxy voting and escalation of concerns during shareholders' meeting. Berk and van Binsbergen (2021) examined the financial impact on ESG divestiture strategies. The paper concluded that such impact is insignificant to have any meaningful change. Instead of divesting, it was recommended for socially or sustainable conscious investors to continue their investment to enact change through stewardship.

An inconclusive answer is made even more complex by how different rating agencies provide inconclusive rating scores of companies and securities. For instance, a security's score by Sustainalytics can differ from its score by MSCI or S&P Global. To harmonise the wide variety of rating standards, the European Commission created a technical standard to categorise them into Article 6, 8 or 9 of the Sustainability Financial Disclosure Regulations (SFDR)—more on that later. By looking at SFDR as a new set of standards, this paper attempts to examine whether sustainable investing costs investment performances or not.

Specifically, investment funds based on the SFDR article standards were chosen for this study. Investment funds reap the diversification benefits through their constituents of many different assets, and they remain popular among individual and institutional investors, especially plan sponsors, such as pension funds who have a matching portfolio for their liability-driven investing. To avoid the smart money hypothesis, mutual funds (SICAVs), both open- and close-ended types, were not chosen. Any outperformance of (non-)sustainable rated funds could be due to the portfolio manager's skill and not its sustainability. While extensive research is poured into studying the performances during the crisis of sustainable equities and mutual funds, few can be found to have investigated exchange-traded funds—even scarce during the COVID-19 crisis.

ETFs are directly investible and flexible as they can be traded throughout the day, have better information transparency, lower costs than open-ended funds, and, specifically for retail investors, tax efficiencies as well. ETFs combines the flexibility of listed stocks and diversification of mutual funds. Furthermore, studying ETFs adds to the utility of its output for investors. As such, exchange-traded funds (ETFs) of SFDR articles 6, 8, and 9 were chosen to examine for performance differences.

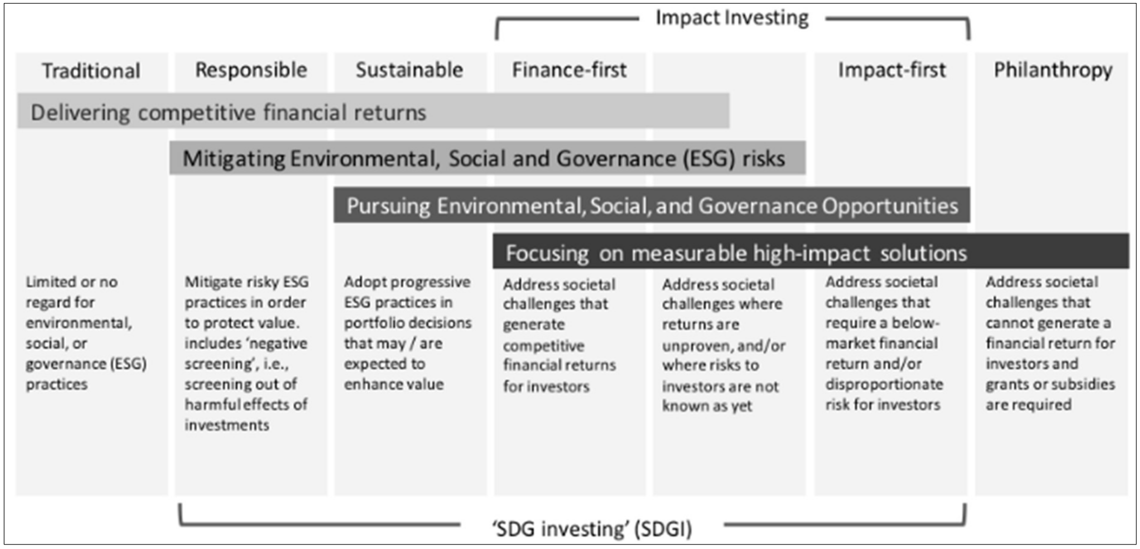
Consequentially, this paper will provide a general overview of sustainable investing, the background of SFDR, why it is an important area for research, what is original for this study, how this paper adds to the literature, the objectives of research, and, consequentially, the relevance of outcome to institutional and individual investors.

1.1 General Overview of Sustainable Investing

To what extent do sustainable investing activities encompass the entire investment world? A report on the sustainable and impact investing trends in the United States (US) market found that one-third of the US assets under management (AUM) are tied to sustainable investing strategies (US SIF Foundation, 2020). Specifically, the assets invested domestically—through sustainable investing strategies—stood at US$17.1 trillion at the end of 2019, a 42% increase from two years prior. The sustainable investing strategies reported by US SIF Foundation include strategies that are ESG incorporated and active ownership, such as shareholder engagement through proxy voting on ESG issues.

This paper categorises the many sustainable investment strategies into the following: ESG investing, responsible investing, impact (finance-first) investing, and impact (impact-first) investing.

Figure 1. Types of Sustainable Investing Strategies


Impact Investing
Traditional
Responsible
Sustainable
Finance-first
Impact-first
Philanthropy
Delivering competitive financial returns
Mitigating Environmental, Social and Governance (ESG) risks
Pursuing Environmental, Social, and Governance Opportunities
Focusing on measurable high-impact solutions
Limited or no regard for environmental, social, or governance (ESG) practices
Mitigate risky ESG practices in order to protect value. includes 'negative screening', i.e., screening out of harmful effects of investments
Adopt progressive ESG practices in portfolio decisions that may / are expected to enhance value
Address societal challenges that generate competitive financial returns for investors
Address societal challenges where returns are unproven, and/or where risks to investors are not known as yet
Address societal challenges that require a below-market financial return and/or disproportionate risk for investors
Address societal challenges that cannot generate a financial return for investors and grants or subsidies are required
'SDG investing' (SDGI)

Source: C-Change (2017). *SDG Investing: Advancing A New Normal in Global Capital Markets*. Retrieved from https://www.c-change.io/blogs/2017/8/3/time-for-a-new-normal-in-global-capital-markets-advancing-investment-in-the-sustainable-development-goals-sdgs

ESG investing, while not illustrated in figure 1, lies between traditional and responsible investing. Such a strategy considers ESG factors but not integration when making an investment decision—just like any other financial metric. Moving on, responsible investing (also known as socially responsible investing) is represented by the responsible and sustainable investing columns in figure 1. The second strategy category goes one step further by integrating (or incorporating) ESG factors into investment decisions—security selection and portfolio construction—and having the active ownership element, the latter known as stewardship. Impact investing, on the other hand, can be focused first on financial returns or the impact. The investment managers commonly align responsible and impact investing for alignment to the United Nations' 17 Sustainable Development Goals (SDGs)—SDG Investment (SDGI). On a side note, the United Nations (UN) estimates that up to USD7 trillion yearly is needed to invest and achieve these 17 SDGs. Only half the amount is currently being met via public and private investments. A funding gap of approximately USD 3 trillion annually is needed to improve our well-being, the earth and society.

This paper focuses on ESG and responsible investing and will not consider impact investing or philanthropy (impact-only investing). Additionally, ESG incorporated strategies (responsible investing) differ from ESG consideration strategies (ESG investing). In the latter, the investment manager considers ESG factors as part of their investment process, just like traditional financial metrics are considered. No negative (or exclusionary) screens, stewardship, or impact analysis are used. For the former, ESG incorporation (or integration) meant one step up, whereby negative screenings are used to exclude certain companies. The ESG factors are integrated into the security selection and portfolio construction process. From this point on, sustainable investing and responsible investing will be used interchangeably.

1.2 Sustainable Financial Disclosure Regulation (SFDR)

The lack of cohesion in the industry when differentiating the terminologies under sustainable investing has also created widespread confusion. For instance, socially responsible investing (or responsible investing for short), ESG investing, and impact investing tend to be used interchangeably—they are not the same. A cohesive international taxonomy to unify the classification of investments, based on the level of sustainability and impact of investments, is missing too and only further adds to the confusion. A prime example can be taken from the different ESG scores a company can have under different ESG rating agencies. Investment into the securities of a listed company may have a high ESG score in Sustainalytics but a lower ESG score in MSCI, creating a lack of consistency and reliability for investment managers during their securities selection process.

Fortunately, at the turn of this decade, the European Commission, the executive body of the European Union, has set about harmonising and resolving the issue. The Sustainable Financial Disclosure Regulation (SFDR) categorises the sustainability profile of investment funds into articles 9, 8 and 6. The categorisation through SFDR has been in effect since March 2021. The SFDR forms part of the EU's Sustainable Finance Framework, which includes the EU Taxonomy.

1.2.1 SFDR Article 9 funds (sustainable investment products)

Funds that have sustainable goals as their objectives are allowed by investment managers or financial institutions to classify them as article 9. Three requirements must be met:

1. The product/funds contribute to a social and/or environmental goal.
2. The investee (invested companies) has good governance.
3. The investment impact does not harm another sustainability objective (such as the UN 17 SDGs).

Conferring to figure 1 type of sustainable development goal investing (SDGI), investments into article 9 funds/products will come under the impact investing strategy.

1.2.2 SFDR Article 8 funds (ESG-integrated products)

Article 8 funds are a step down from article 9 funds, whereby the investments account for ESG factors and do not have E or S as their investment objectives. Like article 9 funds, negative screenings are in place, and investees are required to have good governance. Referring to C-Change's classification of strategies, investments into article 8 funds are under responsible investing strategy. However, while some financial institutions, such as ABN AMRO, considers article 8 as ESG investment instead of responsible investment, ESG investment only considers ESG factors but not integrate them. Looking at the definition of article 8 funds, which has negative screening, ESG investment is, therefore, the wrong terminology used by ABN AMRO.

1.2.3 SFDR Article 6 funds (all other funds; non-sustainable products)

Funds that do not include any sustainability aspects in their investment process. For instance, due to the lack of negative screening, such funds may include investments in fossil fuel companies or tobacco companies. While some products/funds are considered for any sustainability risk, consideration of the risk is insufficient for the funds to be sustainable.

1.3 Area of research, its originality, the relevance, and its objective

Having delved into what responsible investing is and how it differs from impact, ESG or traditional investing, this paper is motivated to explore the difference in the performance of articles 8 and 9 funds against article 6 (non-sustainable funds). To do so, ETF is chosen as the asset class for this study. Additionally, to understand whether responsible investing strategies were resilient throughout the COVID-19 pandemic as an exogenous shock, the performances of these three types of funds will be regressed over the COVID-19 impact that occurred on February 20, 2020. The ETFs investigated will be sorted into two equal-weighted portfolios: a clean (sustainable) portfolio of articles 8 and 9 ETFs and a dirty (non-sustainable) portfolio of article 6 ETFs.

As such, the **main research question** of this paper is the following:

What is the performance difference between a clean and a dirty portfolio, and how does the resilience differ during COVID-19 exogenous shock?

Based on historical analysis of past crises, financial crises (endogenous shock) tend to play out longer when compared to financial turmoil caused by exogenous shocks, such as an epidemic like SARS. Exogenous shocks in the financial markets tend to be short-lived, but the adverse impact is more profound. Guo & Zhou (2021) found that the stock market in China (the epicentre of SARS) increased by 20% during 2003, despite SARS. However, the exogenous shock caused by Covid-19 has been prolonged for over two years due to new variants and lockdown measures. In the first stage of the COVID-19 pandemic, March 2020, stock markets globally found 30% of their value being wiped out. That value has yet to be recovered after one year (Guo & Zhou, 2021).

Consequentially, investors' flight to safe-haven investment assets and hedging instruments to hedge tail risks from such rare shocks are seen. Since 2021, academic research gained momentum regarding the resiliency of a variety of asset classes during exogenous shocks. Combined with the growing trend of responsible investment with ESG-integrated portfolio, studies into performance resiliency and performance difference of listed equities with ESG-screened factors and green bonds have been conducted (Albuquerque et al., 2020; Guo & Zhou, 2021). Guo & Zhou (2021) found that green bonds (a sustainable investment with sustainable objective) are resilient against financial shocks—endogenous shock—but evidence of its resilience against Covid-19 are scanty in the world of academic research.

Therefore, this paper's research contributes to the growing literature on the performance difference of asset classes due to higher ESG scores. It also further contributes to the performance resilience of asset classes with higher ESG scores. Specifically, to the empirical understanding of the opportunity of articles 8 and 9 exchange-traded funds, it may present itself as a safe-haven asset against tail risks. The study is original by addressing the old problem—different ESG scores from different ratings—with a new way through the SFDR lens. Frequently, academic research has had to grapple with multiple, conflicting ESG scores from rating agencies.

Apart from the academic relevance, the professional usefulness of the research for professional and institutional investors is to conclude whether sustainable investing presents itself as a massive business opportunity. The opportunity of not just reaping financial rewards but also contributing to the shared conquest of environmental and societal challenges present in our times.

Following on from chapter 1, chapter 2 will continue with a review of the literature concerning investments into asset classes with higher ESG scores and their performance against those with lower ESG scores. Moreover, whether investments into asset classes with higher ESG scores will be more resilient (against exogenous shock) than those of lower ESG score. By resilience, this paper will look at whether returns are relatively higher than benchmark indices and those of lower ESG scores. The volatility, higher order moments (kurtosis and skewness), and factor sensitivities from multifactor models will also be examined. Lastly, the development of four hypotheses to address the main research question will be elaborated on in detail. The use of conceptual frameworks will be illustrated in chapter 3 for the readers.

The dataset's construction will be discussed in chapter 4, including the data source and data collection process. Chapter 4 will end with post-estimation statistics to check on the data against assumptions of the classical linear regression model, BLUE[1], and summary statistics of the factor loadings. Chapter 5 follows through with an introduction to the study's methodology and the four methods used to address the four hypotheses of this paper's central research question. In chapter 6, the results of the four methods will be shown, including both the summary statistics and regression results. Moving on to Chapter 7, a discussion will take place to bring all the results together and discuss how they support and oppose other relevant works of literature findings. Chapter 8 will close off with a conclusion, research limitations and recommendations for further research.

[1] Best Linear Unbiased Estimator from Gauss–Markov theorem (Plackett, 1949)

2. Literature review

In this chapter, a review of existing works of literature concerning performance differences and performance resilience will be conducted. Next, four hypotheses will be developed to tackle the main research question. A literature review of methods to test those four hypotheses, such as multifactor models by Fama and French (2015), will be elaborated on below. Appendix A shows an overview of literature finding evidence of sustainability performance and resilience. Appendix B shows an overview of literature finding no evidence or even evidence against sustainability performance and resilience.

2.1 Performance difference of investments with higher ESG

Various empirical studies have researched the performance difference of investments into products with a higher ESG score against those with a lower ESG score. However, scholars have little consensus regarding the significance of performance differences with responsible investing. To start with, the most persistent belief against responsible investing concerns underperformance. This basis lies behind Markowitz (1959) Modern Portfolio Theory, whereby the risk-return relationship and the efficient frontier of a portfolio were illustrated. Using negative screenings based on ESG factors, an investor limits the investment universe due to non-financial reasons and inadvertently increases the risk of underperformance and tracking error. Not only are the investment opportunities constrained, but the diversification efficiencies are reduced as well (Lee et al., 2010).

Most of the prior decades' research pointed towards no difference between one with higher ESG against one with lower ESG. Early research on the period before this century's start concurs with the finding. In an early study of ethical and conventional funds between 1990 and 2000, Bauer et al. (2005) concluded no significance in financial performance differences in risk-adjusted returns (Jensen's alpha). The ethical funds underwent a catching-up phase before generating similar risk-adjusted returns to comparable conventional funds. Brammer et al. (2006) went one step further by concluding that responsible investing can even hurt financial performances and that investing in conventional funds and companies is required for maximised returns. However, Brammer et al. (2006) only examined one single period based on variables observed on July 1, 2022. When zooming into US sustainable and responsible funds market, Girard et al. (2007) looked at over 100 open-ended funds between 1984 and 2003. The authors concluded they underperformed conventional funds by 7% to 9% annually on average.

Moving on to more recent research, results also led to similar conclusions. While Lee et al. (2010) found that increased screening intensity on the investment universe lowers systematic risk—it does not affect idiosyncratic risk—it reduces overall risk due to lower beta for stocks. There is a reduction of 70 basis points per screen applied by the investment manager. Investors also pay the price for responsible investing through socially responsible investment (SRI) funds. SRI funds were found to have underperformed against their respective benchmarks globally by -2.2% to -6.5% (Renneboog et al., 2008). In that same study, using screens for social and/or governance factors also led to lower risk-adjusted returns. As for the European market, Cortez et al. (2012) found that. US and Austrian SRI funds even showed evidence of underperformance.

Apart from research into specific asset class performance—such as the mutual funds, listed equities, and bonds—empirical evidence into the portfolio of higher-ESG asset classes also proven inconclusive regarding performance differences. Constructing a portfolio with companies having high ESG factors was found to have not paid off by some research, while it does pay off by other research studies. The abnormal returns from a negative screening strategy were not statistically significant compared to a strategy of not screening for ESG factors (Halbritter & Dorleitner, 2015; Auer & Schuhmacher, 2016). In other words, these two papers concluded that responsible investing did not deliver superior risk-adjusted returns (Jensen's alpha). In fact, like the finding of Lee et al. (2010) and Renneboog et al., (2008), exclusionary screenings do lead to substantial costs for investors (Adler & Kritzman, 2008).

On the flip side, positive screenings were found to have yielded results. The use of positive screens to discover high ESG score companies, compared to narrowing down the investment universe with negative screens, was concluded to have delivered statistically significant abnormal returns for portfolios between the period 1992 to 2004 by Kempf & Osthoff (2007) and between the period of 1992 to 2007 by Statman & Glushkov (2009). To bring in a non-academic context for insights into what the industrial reports think about investments with higher ESG, a review of several professional journals and industrial reports was conducted.

A recent report on sustainable funds' outperformance in 2020 from Morningstar's sustainability matter collections stood out the most. On active management, sustainable equity funds were found to have outperformed their conventional peers. In that report, Hale (2021) sorted sustainable and traditional equity funds into four quartiles based on their returns in 2020. 42% of the sustainable equity funds were represented in the top quartile, while 6% were in the bottom quartile. Hale (2021) found that these sustainable equity funds were exposed to quality and growth stocks when looking at factor tilts. While growth stocks relate to the value factor, quality stocks have two different meanings. MSCI defines *quality factors* as stocks with high profitability, stable earnings, and low leverage. Fama and French (2015) define quality with two factors: investment policy and profitability.

Apart from sustainable equity funds' outperformance relative to their conventional peers, the report also concluded that sustainable index funds (passive management) have outperformance in funds inflows—the inflows were double of actively managed sustainable funds. A point of interest comes from the underperformance of these sustainable index funds in developed markets outside the United States and Canada. Only around one-quarter of them outperformed the MSCI EAFE Index. As a result, this paper narrows the research into ETFs from Europe to provide empirical research findings that are not US market focused.

On market beta, Lagewaard (2020) found that when socially responsible funds do not invest in any controversial industry, it results in a lower market beta compared to conventional funds. Conventional funds were also found to lower excess returns when looking at Jensen's alpha of risk-adjusted returns.

Consequentially, it is interesting to examine the new designation of exchange-traded funds with SFDR article 8 (light green funds) and article 9 (dark green funds) and compare them against article 6 (non-sustainable funds). The amount of article 9 ETFs is currently much lesser in the investible universe than article 8 ETFs. Many institutional investors also consider responsible investing strategy and not impact investing strategy at the current moment, which is why the clean portfolio consists of articles 8 and 9 ETFs. This leads to the following first hypothesis:

H1: There is no statistical significance in performance difference between a clean and dirty portfolio

The first hypothesis belongs to part 1 analysis using monthly returns and regressing it using multifactor models. A conceptual framework for part 1 analysis can be seen below in figures 2 and 3 of chapter 3 conceptual framework. The part 1 analysis uses Fama and French (1993) 3-factor models, which expand on the capital asset pricing model (CAPM) by adding size and value factors. The expected outcome of this paper is to reject the first hypothesis and conclude that responsible investing does bring in a statistically significant performance difference. In a study of risk and returns of sustainable funds between 2004 and 2018, Morgan Stanley Institute for Sustainable Investing (2019) found no significance in total returns difference between ESG-focused and traditional mutual funds and index funds—no financial trade-offs between the two. Nevertheless, when looking at risk, the sustainable funds had a lower market risk as they had 20% lesser downside deviation than traditional funds. This brings the research to the second hypothesis:

H2: A clean portfolio performed better in these metrics using performance measures than a dirty portfolio.

Using performance measures to rank and evaluate performances is common in portfolio evaluations. Often, investment returns do not follow a normal distribution curve and are asymmetric. Markowitz (1959) recommended using semi-variance as an alternative measure by considering risk-adjusted returns from the negative deviation point of view. Certain investment returns are even more asymmetric when investments are directed towards asset classes with higher tail risk. Hedge funds are a prominent example of the widespread use of performance measures to evaluate their returns using non-conventional measures such as the Sortino ratio.

The two most popular measures are the Sharpe ratio and Jensen's alpha. However, these two measures come with their drawbacks. As the ETFs are tracking indices, the Sharpe ratio computed is sensitive to the market index (Roll, 1978). Furthermore, Dybvig and Ross (1985) found that Jensen's alpha is influenced by market timing. Market timing from a fund manager can lead to biased estimates for Jensen's alpha. Hence, several performance measures will examine the second hypothesis, expecting the results to align with the second hypothesis.

Looking through the lens of sustainable indices, the long-run outperformance of sustainable indices relative to the S&P 500 is in the majority. A prime example can be taken from the MSCI KLD 400 index, the oldest sustainable index globally, which outperformed the S&P 500 by 81 basis points (annualized) from 1990 to 2016.

If a sustainable fund has no underperformance relative to conventional funds, it is good news, too, even if there is no statistical significance in performance differences. This is because these sustainable funds (looking at just ESG-integrated and not impact funds) are satisfying the sustainability needs of their investors while not sacrificing financial returns.

2.2 Resiliency of article 8 funds throughout COVID-19 pandemic as an exogenous shock

Responsible investing through articles 8 and 9 ETFs may not be resilient over the exogenous shock. At the time of writing, in the second quarter of 2022, sustainable equity indexes faced headwinds from the first quarter of 2022. Since the Russian invasion of Ukraine—another exogenous shock—the financial markets and global economy have been in turmoil. Energy stock rally and technological stock sell-off, due to disappointing earnings results from firms—such as Meta—and rising inflationary rates were two significant trends of Q1 2022.

Morningstar US Sustainability Leaders Index—tracking 50 of the best sustainability score of US large caps—ended the first quarter of 2022 with -10.56%. Morningstar US Sustainability index, another sustainability equity index, was down -6.90% in Q1 2022. However, on a longer-term basis, sustainability seems to help systematically drive investment performances relative to their traditional peers. As of 31 March 2022, Morningstar's US Sustainability Leaders Index had 17.30% and 19.98% absolute returns on a 5-year and 3-year horizon. For Morningstar's US Sustainability index, it was 15.52% and 18.03% absolute returns on the same 5-year and 3-year horizon.

When looking at specific asset classes, Albuquerque et al. (2020) found that stocks with higher environment or social ratings have lower volatility, higher returns, and higher operating profitability during Q1 2020—the first wave of the COVID-19 pandemic. In terms of governance factors, firms engaging in corporate social responsibility (CSR) towards societal and environmental contributions were found to benefit shareholder value creation (and preservation during exogenous shocks) and their returns (Ferrell et al., 2014; Servaes & Tamayo, 2013; Henisz et al., 2013; Hillman & Keim, 2001). CSR practices can also make firms resilient against societal protest as an exogenous shock. In the 1999

Seattle World Trade Organisation protest, Schnietz & Epstein (2005) found that firms which have a reputation for CSR were protected from market shocks following the protest.

Moving on from exogenous shock to focusing on endogenous shock resilience, there exists literature supporting resilience in times of financial crisis too. Lins et al. (2017) concluded that companies with higher CSR outperformed competitors with lower CSR levels during the global financial crisis. Looking through the lens of mutual funds instead of listed equities, Nofsinger & Varma (2014) concur that responsible mutual funds outperformed conventional funds in terms of risks. Those lower number of lawsuits, better reputation, and stable relationships that the investees have with the responsible mutual funds were drivers against bankruptcy risks during a depressed market.

Currently, limited research papers are available on the performance of SFDR funds during adverse market conditions caused by COVID-19 exogenous shock. Hence, it is interesting to investigate whether articles 8 and 9 EFTs are more resilient in times of such exogenous shocks and put any ambiguity of responsible funds' risk-mitigating benefits to rest. Hence, the third hypothesis is:

H3: There is no statistical significance in terms of factor loading sensitivities for a clean portfolio during COVID-19.

Fama and French (2015) 5-factor models and Carhart (1997) 4-factor models will be used alongside CAPM and the 3-factor model. A 6-factor model will also be regressed using the momentum factor from Carhart (1997) and the 5-factor model. The expectation of the outcome will be the rejection of the third hypothesis and that responsible investments through a clean portfolio are indeed resilient during COVID-19. When looking at factor sensitivities, there is statistical significance for a clean portfolio, but they are less sensitive than a dirty portfolio. Furthermore, a clean portfolio is expected to have statistically significant outperformance relative to a dirty portfolio in alpha. The relevance of the outcome will imply whether responsible investments are beneficial against future virus shocks the investments may face.

Regarding sustainability strategies, through responsible investing (article 8 funds) or impact investing (article 9 funds), they tend to have technological stock tilts while underweighting energy stocks—they carry higher ESG risks.

Lastly, a macroeconomic factor model will be used to investigate the resilience of clean portfolios. Studies into the significance of macroeconomic variables' influence on asset classes have been inconclusive. Still, most research has shown evidence of significance in the relationship between stock returns and macroeconomic factors (Tangjitprom, 2012). On the one hand, when it comes to the influence of macroeconomic variables on US stock returns, Çiftçi (2014) found no significance in the influence between interest rates on stock returns and exchange rates on stock returns. In that paper, macroeconomic variables had largely no impact on stock returns, except for crude oil and gold—their influence differs between different sectors' stocks.

On the other hand, Flannery and Protopapadakis (2002) conclude that some macroeconomic variables impact US stocks. The consumer price index (inflation) is significantly correlated with stock returns, while the employment report and balance of trade factors only affect the US stocks' volatility. Inflation, especially during a high inflationary environment, has an inverse relationship with asset classes (Kolluri & Wahab, 2008). A high inflation rate erodes the real value of an investment, such as an interest-bearing mutual fund.

The state of the economy influences macroeconomic factors and impacts the financial market too. For instance, Boyd et al. (2005) concluded that an increasing unemployment rate positively influences stocks in an expansionary economy while negatively influencing stocks during a contractionary economy. Hence, this paper examines whether the deterioration of key macroeconomic indicators—during COVID-19—will influence the performance of a clean and a dirty portfolio. This brings the paper to the fourth and last hypothesis::

H4: There is no statistical significance in terms of macroeconomic factor loading sensitivities for a clean portfolio.

The outcome will expect to reject the fourth hypothesis and find that at least one of the macroeconomic factors used will be statistically significant. Moreover, comparing the clean and dirty portfolio's coefficient estimates, the clean portfolio will be less sensitive than the dirty portfolio in terms of the macroeconomic factors as regressors.

3. Conceptual Framework

Conferring from the section above, the analysis will be a two-pronged approach: segment 1 on performance difference and segment 2 on performance resilience against COVID-19 exogenous shock. In segment 1 (see section 5.1), the analysis will be sub-divided into part 1, using CAPM and multifactor analysis, and part 2—using performance measures.

For segment 2 (see section 5.2), using an expanded multifactor analysis—part 3—and macroeconomic factor analysis (part 4) will shed light on how resilient articles 8 and 9 ETFs during COVID-19 are and how they fare relatively against article 6 ETFs. Apart from the performance measures, three conceptual frameworks are illustrated below to highlight the model specifications that will be time-series regressed upon.

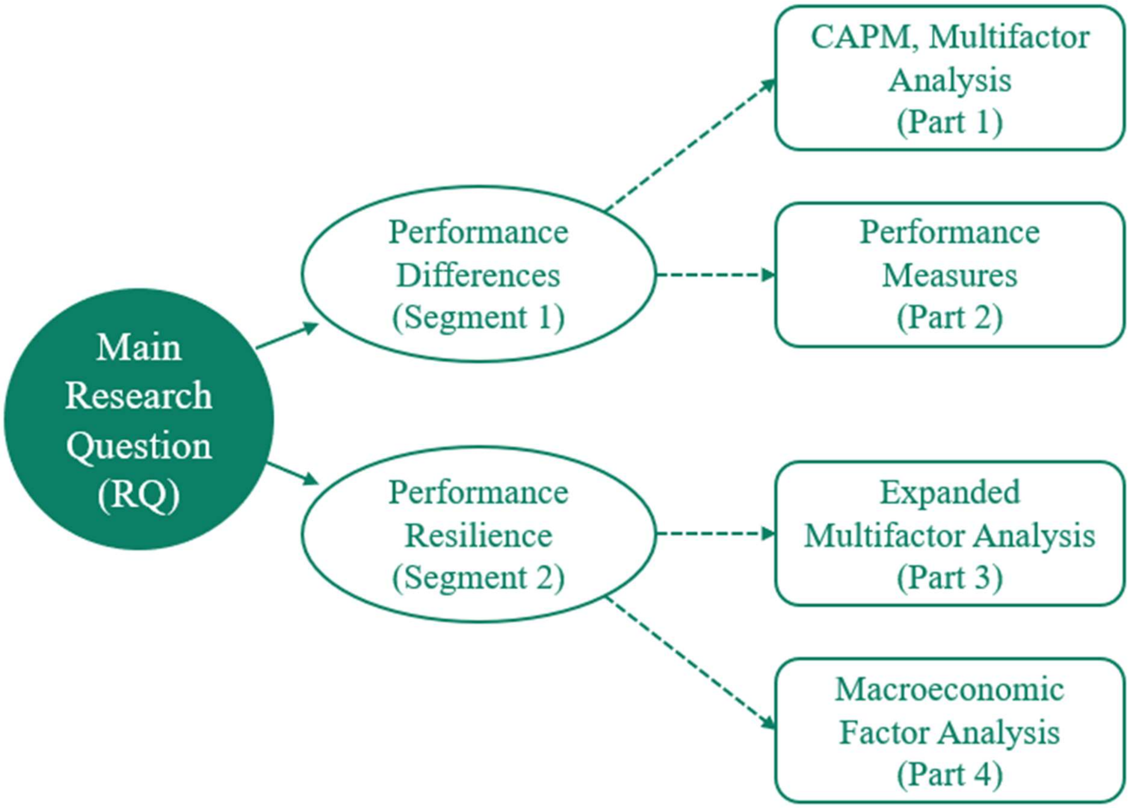

Main Research Question (RQ)
Performance Differences (Segment 1)
Performance Resilience (Segment 2)
CAPM, Multifactor Analysis (Part 1)
Performance Measures (Part 2)
Expanded Multifactor Analysis (Part 3)
Macroeconomic Factor Analysis (Part 4)

Figure 2. Conceptual Overview of Analysis

In part 1 (figure 3), the capital asset pricing model (CAPM), also known as the single factor model, and Fama-French 3 (FF3) factor model, will be used to analyse the returns' performances. The comparison will be conducted between the clean portfolio (article 8 and 9 ETFs) and the dirty portfolio (article 6 ETFs). The two portfolio constructions are equal-weighted, and their respective monthly total returns are from the Morningstar database. The period is from January 2017 to May 2022, giving 65 monthly return observations. Time-series regression with Newey-West standard error was conducted on the single and multifactor models. 1-month Euribor was chosen as the risk-free rate (Rf), while the

MSCI Europe IMI Index's returns will serve as the market rate (Rmt). MSCI AC Europe Small Cap Index and MSCI AC Europe Large Cap Index were used to proxy for the size factor. Lastly, the indices—MSCI Europe IMI Growth and MSCI Europe IMI Value—were also used to proxy for the value factor.

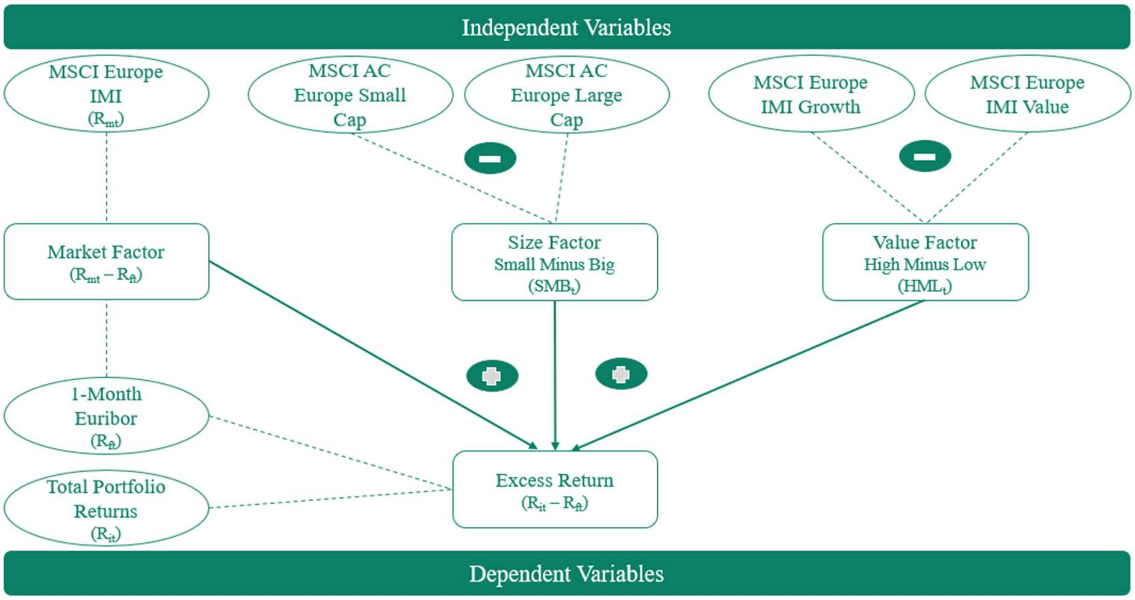

Independent Variables
MSCI Europe IMI (R_{mt})
MSCI AC Europe Small Cap
MSCI AC Europe Large Cap
MSCI Europe IMI Growth
MSCI Europe IMI Value
Market Factor ($R_{mt} - R_{ft}$)
Size Factor Small Minus Big (SMB_t)
Value Factor High Minus Low (HML_t)
1-Month Euribor (R_{ft})
Total Portfolio Returns (R_{it})
Excess Return ($R_{it} - R_{ft}$)
Dependent Variables

Figure 3. Conceptual Framework of Part 1 Multifactor Analysis (Monthly Returns; Indices as Factor Proxies)

In part 3 analysis (figure 4), the multifactor analysis will now go from 3 to 6 factors. They now include the momentum factor from Carhart (1997) and two additional quality factors—profitability and investment policy—from Fama and French (2015), which were expanded upon from their three-factor paper back in 1993. Part 3 differs from part 1, with daily returns being used, instead of monthly returns. The factor loadings data were also taken from Professor French's data library from Dartmouth College[2], compared to using indices as factor proxies in part 1. Moreover, although the factor loadings are for Europe, the risk-free rate and currency denomination differ. Three different datasets from the data library were taken: Fama-French 3 factor loadings (market, size, value); Momentum factor loadings; and Fama-French 5 factor loadings (market, size, value, profitability, investment policy). As the risk-free rate is a 1-month US treasury bill rate and the currency is US Dollar denominated, the returns were converted to US dollar at the day's closing spot rate.

[2] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

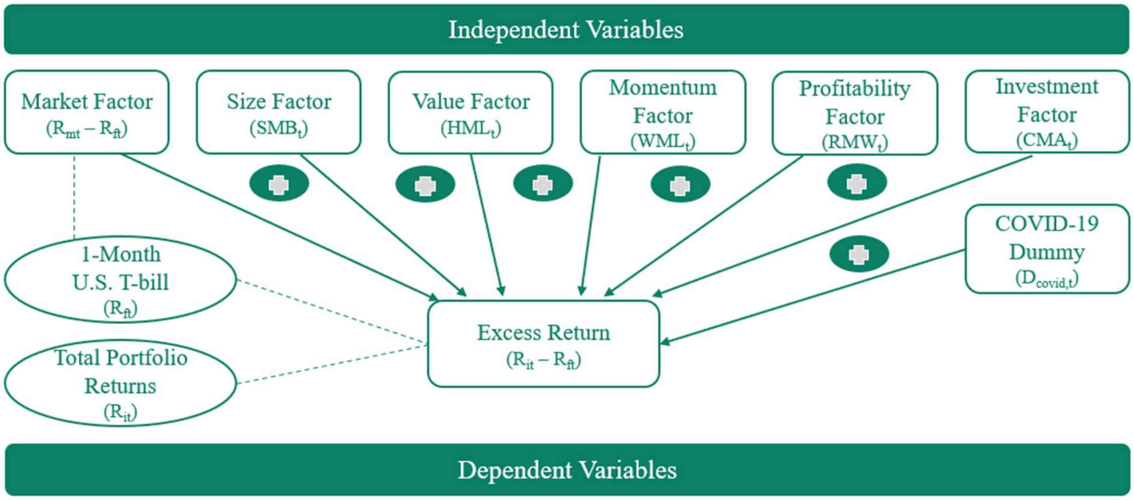

Independent Variables
Market Factor $(R_{mt} - R_{ft})$
Size Factor (SMB_t)
Value Factor (HML_t)
Momentum Factor (WML_t)
Profitability Factor (RMW_t)
Investment Factor (CMA_t)
1-Month U.S. T-bill (R_{ft})
COVID-19 Dummy $(D_{covid,t})$
Excess Return $(R_{it} - R_{ft})$
Total Portfolio Returns (R_{it})
Dependent Variables

Figure 4. Conceptual Framework of Part 3 Expanded Multifactor Analysis (Daily Returns)

In sub-section 5.2.1 from chapter 5, five different model specifications were regressed. For CAPM single factor model, Fama-French 3-factor model and Carhart 4-factor model, the 3-factor loading dataset and momentum factor loading dataset from Prof. French's data library were used. For the Fama-French 5-factor model and Fama-French 5+momentum factor model, the 5-factor loadings dataset was used instead, alongside the same momentum factor loading dataset.

To examine for the resilience of clean portfolio against COVID-19, a dummy variable is added to codify the daily returns into two periods (0 for before the COVID-19 recession of 19 February 2020 and 1 after). The period studied is 180 days before COVID-19 (5 June 2019 to 19 February 2020) and 180 days during COVID-19 (20 February 2019 to 5 November 2020). In total, 360 daily returns as observations.

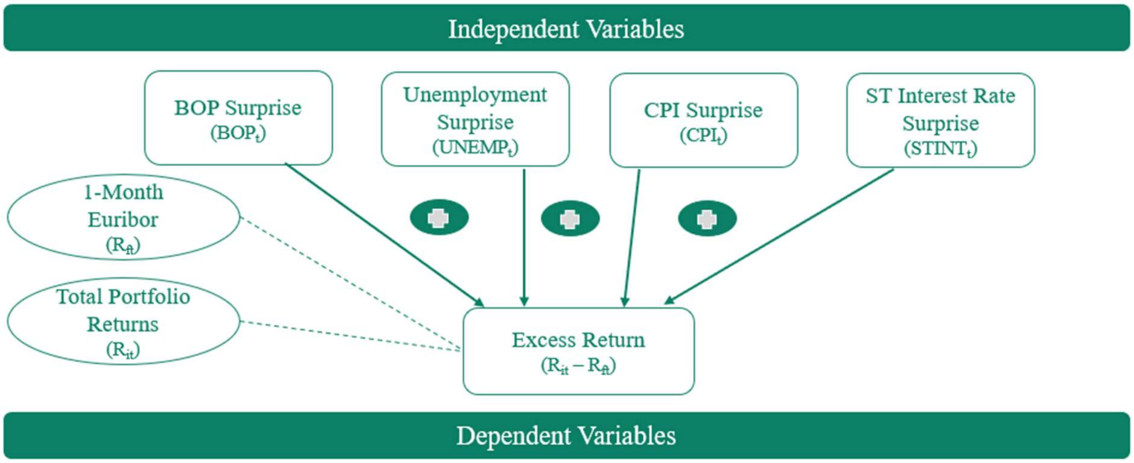


Figure 5. Conceptual Framework of Part 4 Macroeconomic Factor Analysis (Quarterly Returns)

Lastly, figure 5 depicts a macroeconomic factor model to analyse the impact on the returns of the portfolios. These factors illustrate the relational link between regional indicators (Euro area) and the performances of financial markets (represented by the ETFs) and how the deterioration of these factors due to COVID-19 can impact the portfolio returns—or lack thereof. Here, the deterioration of the macroeconomic indicators serves as a proxy for COVID-19 as an exogenous shock instead of using a dummy variable.

Suppose there is a significant relationship between (one of) the factors and the returns. In that case, analysis can examine whether the clean portfolio is more resilient than the dirty portfolio. Furthermore, insights can be derived on which macroeconomic indicators are more intricately connected with the portfolio performances.

Tangjitprom (2012) classifies the macroeconomic factors into four categories:

1. *General Economic Conditions*: Industrial Production; Gross Domestic Savings, Consumption; and Employment Level
2. *Interest Rate and Monetary Policy*: Interest Rate; Term Spread; Default Spread; and Money Supply
3. *Price Level*: Consumer Price Index; Crude Oil; and Gold
4. *International Activities*: Exchange Rate; Foreign Direct Investment; and Foreign Exchange Reserve

From each category, one macroeconomic factor was chosen, and they are computed for their surprise ratios (difference between actuals and expected figures). The four factors are: Unemployment rate (UNEMP$_t$), Consumer Price Index (CPI$_t$), Short-Term Interest Rate (STINT$_t$), and Balance of Payment Current Account (BOP$_t$). BOP chosen as the international activity for Euro area.

4. Data

In this chapter, the data collection will be elaborated on in detail, including the source, process, and construction of datasets for the time-series regression. Three sets of return frequency (daily, monthly, quarterly), resulting in 6 different portfolios (clean and dirty portfolios), were constructed. They are differentiated by their frequencies of returns (daily, monthly, quarterly) and SFDR articles (article 8 and 9 for clean portfolio, article 6 for dirty portfolio).

4.1 Data Collection

First, a selection of ETFs must be made before obtaining the data for the dependent variable (the returns of each ETF). To do so, the fund's prospectus and list of funds from major asset managers in Europe were examined. This manual undertaking—at the time of writing—was conducted as financial databases, such as Morningstar or Bloomberg, do not offer the option of adding SFDR article types as a criterion in investment universe screening—they offer their sustainability rating as a criterion. The ETFs from Europe, denominated in Euro, are segregated according to their SFDR Article classification.

Table 1. ETF summary (number of ETFs per frequency)

	Daily Returns	Monthly Returns	Quarterly Returns
	5 June 2019 to 5 November 2020	*January 2017 to May 2022*	*Q4 2010 to Q4 2021*
SFDR Articles	*(n = 360 days)*	*(n = 65 months)*	*(n = 45 quarters)*
Article 6	57	51	35
Article 8	25	17	10
Article 9	4	3	2

Table 1 offers an overview of the sample size of ETFs; tables 6, 9 and 11 in chapter 6 shows further descriptive and summary statistics. The number of ETFs officially classified as articles 8 or 9 are much lower than those of article 6. It is in reverse when examining mutual funds (SICAVs in Europe). The number of SICAVs in article 6 is much lower, while in articles 8 and 9 SICAVs are much more. To digress, this can be explained by how a better sustainability rating attracts more funds inflow and—consequentially—their compensation as fund managers, as well as the rise of greenwashing (Bauer et al., 2021; Ben-David et al., 2021; and Del Guercio & Tkac, 2008). On the other hand, ETFs are seeing more creations being classified as articles 8 or 9 since COVID-19 (Gantchev et al., 2020; Bauer et al., 2021). This explains why daily returns, which have the period that is from the recent past, have more article 8 and 9 ETFs. Monthly and Quarterly returns have lesser ETFs of articles 8 and 9 due to lack of data availability and the absence of them prior to COVID-19.

4.1.1 Dependent Variable

The Morningstar database is the source of the data for the dependent variable—total returns—of each ETF extracted (in daily, monthly, and quarterly frequencies). Those total returns were selected as the variable to download from Morningstar because it is already expressed in percentage terms and accounts for the change in price, reinvesting, and distribution. Total returns give a more accurate view for portfolio construction and comparison as it does not adjust for the sales charges (e.g., front-end load), but it does account for the expense ratio.

4.1.2 Independent Variables (Part 1 CAPM, Multifactor Analysis)

Referring to figure 2, indices as proxies for the factor loadings (regressors) were used. For the size and value factor, four different indices were used from MSCI. These data were extracted from MSCI's end-of-day index data search[3], and in monthly frequency from January 2017 until May 2022. To recap, the four indices were: MSCI AC Europe Small Cap Index, MSCI AC Europe Large Cap Index, MSCI Europe IMI Growth Index, and MSCI Europe IMI Value Index. To obtain the size factor, the relative returns movement of AC Europe Small Cap Index was deducted against AC Europe Large Cap Index. In the same vein as the size factor, the value factor was determined by MSCI Europe IMI Growth Index and MSCI Europe IMI Value Index.

A fifth index was used for the market factor loading—MSCI Europe IMI Index. IMI, which stands for investable market index, captures approximately 99% of the free float-adjusted market capitalisation of large, medium, and small capitalisation constituents. There are 1,475 constituents in the index, spanning across 15[4] developed markets of Europe. Thus, the index is a good market proxy for Europe.

Finally, on the risk-free rate, 1-month Euribor was chosen and extracted from the European Central Bank (ECB) Statistical Data Warehouse[5].

4.1.3 Independent Variables (Part 3 Expanded Multifactor analysis)

Part 3 of the conceptual framework (see figure 3) forms one of the two analyses towards examining how resilient the clean portfolio is against COVID-19 and relative to a dirty portfolio. The total daily returns are used as the regressand compared to monthly total returns for part 1. The 6-factor loadings—market, size, value, momentum, investment policy and profitability—come from Prof. French's data library[6] and courtesy of Dartmouth College. The factor loadings are specific to Europe,

[3] https://www.msci.com/end-of-day-data-search

[4] Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the UK.

[5] https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=143.FM.M.U2.EUR.RT.MM.EURIBOR1MD_.HSTA

[6] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

but the factor loadings were based on US dollars, and the risk-free rate is the US 1-month T-bill rate. Hence, the daily returns in the regressand were converted to US dollars from Euro, based on the day's closing exchange rate.

Without digressing further, Prof. French's construction for the factor loadings will be described briefly. On size factor, it is the mean returns differences of his 9 (3x3) portfolios between small and big portfolios. On the value factor, it is the average return differences between 2 value portfolios and 2 growth portfolio stocks. The two quality factors—profitability and investment policy—are also constructed similarly to the fundamental factors (size and value). The profitability factor was constructed by taking the average returns of 2 robust portfolios and deducting off the average returns of 2 weak portfolios. The investment policy factor was constructed similarly using 2 portfolios of stocks with conservative investment policy and 2 with aggressive policy. Lastly, the momentum factor is the average returns of 2 winner portfolios for Europe minus the average returns of 2 loser portfolios for Europe.

4.1.4 Independent Variables (Part 4 Macroeconomic Factor Analysis)

In part 4, see figure 4, macroeconomic indicators were used this time as the factor loadings. In addition, quarterly frequency is utilised, as is the norm for macroeconomic factor analysis. 1-month Euribor (quarterly frequency) is the risk-free rate. Four macroeconomic indicators were shortlisted—explanations on why these four can be found in section 5.2.2—and not more to avoid the implication of overfitting the model. The four macroeconomic indicators are Balance of Payment Current Account (BOP), Unemployment Rate (UNEMP), Consumer Price Index (CPI) and Short-Term Interest Rates (STINT). In the model specification, the surprise ratio of each regressor was used. To obtain the surprise ratio, the actuals and forecasted figures for each indicator were extracted from Organisation for Economic Co-operation and Development (OECD) Data Warehouse[7]. The relative rate of change (%) for each indicator was used, and the indicators were for Euro Area countries.

The unemployment rate serves as a proxy for labour market indicator, while the CPI—including food and energy—serves as the price measure for inflation. BOP is a proxy for international trade indicators, while the short-term interest rate is a proxy for the interest rate environment. These four leading indicators provide insights into how their deterioration (due to Covid) can influence the financial market (or no significant influence). It can also allow an examination into whether a clean portfolio is more resilient when compared to a dirty portfolio.

[7] https://data.oecd.org/

4.2 Dataset Preliminary Checks (Postestimation Statistics)

This section checks the assumptions of residuals of the Ordinary Least Squares (OLS) method. Part 1, 3 and 4 model specifications were regressed using OLS time-series regression with the Newey-West estimator. The assumptions on the error terms (u_t) and the tests and methods to ensure validity are elaborated below. Table 2 shows the results of the post-estimation statistics from the relevant tests below.

Table 2. Post-estimation statistics

		Daily (FF3+MOM Factors)		Daily (FF5+MOM Factors)		Monthly (FF3)		Quarterly (Macroecon)	
Panel A	*Breusch-Pagan*	*Dirty (1)*	*Clean (2)*	*Dirty (3)*	*Clean (4)*	*Dirty (5)*	*Clean (6)*	*Dirty (7)*	*Clean (8)*
	Chi2	8.51	12.94	65.35	62.41	1.05	0.61	3.07	1.16
	Prob > chi2	<.01	<.01	<.01	<.01	.31	.44	.08	.28
	White								
	Chi2	72.87	93.53	170.18	159.17	5.37	6.11	8.81	7.71
	Prob > chi2	<.01	<.01	<.01	<.01	.80	.73	.84	.94
Panel A	*Durbin-Watson*								
	DW d-stats	1.47	1.51	1.75	1.72	0.00	0.00	0.00	0.00
	Breusch-Godfrey								
	Chi2	0.28	0.06	5.04	4.33	0.00	0.00	0.00	0.00
	Prob > chi2	.60	.81	.02	.04	1.00	1.00	1.00	1.00
Panel C	*Skewness-Kurtosis*								
	Adj chi2	7.61	7.02	22.47	18.56	4.31	2.14	10.78	11.48
	Prob > chi2	.02	.03	<.01	<.01	.12	.34	.01	<.01

Note. Column (1) to (4) relate to part 3 of this paper's analysis. Column (5) and (6) relate to part 1 of this paper's analysis, while (7) and (8) relate to part 4. 'Dirty' column stands for dirty portfolio of equally weighted article 6 ETFs. 'Clean' column stands for clean portfolio of equally weighted article 8 and 9 ETFs. FF stands for Fama-French, MOM stands for momentum. Tests are conducted at 5% significance levels.

4.2.1 Assumption 1: Errors have zero mean [$E(u_t) = 0$]

The expected mean of error terms from OLS regression should be zero. i.e., the OLS error terms distribution has a mean equal to zero. To ensure the validity of assumption 1, the constant term is present in each model specification and not omitted. The estimates of R^2 for all models is also non-negative (see chapter 6 results below).

4.2.2 Assumption 2: Homoskedasticity [Var(u_t) = $\sigma^2 < \infty$]

The assumption of variance of errors being constant—homoskedasticity. If the variance is not constant, the model specification will bring incorrect coefficient estimates due to heteroskedastic errors. To overcome this, the Newey-West estimator was used. As the time-series regression was conducted with zero lags (*h=0)*, the estimator is robust against heteroskedasticity but not autocorrelation. Thus, this estimator is, mathematically, the same as Robust Standard Errors, which overcomes heteroskedasticity in error terms and not serial correlation. Panel A of table 2 above shows two tests for heteroskedasticity: The Breusch-Pagan test and the White test.

Breusch and Pagan (1979) developed a test for heteroskedasticity. The method tests for the null hypothesis of homoscedasticity (variances of errors are all equal), while the alternative hypothesis would be heteroskedasticity—variances are not equal. The row *prob>chi2* shows whether the errors are heteroscedastic at *p-value <0.05*. The model specifications, with their resulting errors, under monthly and quarterly columns, have probability values above 0.05—homoscedasticity. However, the probability values for the remaining columns under daily have less than 0.05. This implies heteroscedasticity presence in the residuals (error terms), and the null hypothesis of homoskedasticity of residuals is rejected.

White (1980) also developed a test for heteroskedasticity. Like the Breusch and Pagan test results, the White test's probability values under monthly and quarterly columns are insignificant. The values of the remaining columns are significant. Hence, there is the presence of heteroskedasticity in the residuals, and the estimator should then be changed to one that is heteroscedasticity-consistent standard errors. Failure to do so may lead the estimator to give unbiased, consistent coefficients but no longer the best linear unbiased estimator (BLUE). The standard errors can become misleading as they may be too large for the constant.

As such, the Newey-West estimator on all model specifications was used for all time-series regression.

4.2.3 Assumption 3: No autocorrelation [Cov(u_i.u_j) = 0, for $i \neq j$]

Combining the second assumption of homoscedasticity with assumption 3 of no serial correlation (autocorrelation), the error terms will be independent and identically distributed (i.i.d.). Assumption 3 means errors are linearly independent of one another; the covariance of error terms over time will be zero. If error terms are correlated, they are said to be serially correlated (or autocorrelation). Panel B of table 2 above shows two tests for autocorrelation being employed: Durbin-Watson and Breusch-Godfrey.

Durbin-Watson (1950) test statistics test for the null hypothesis of error terms being not serially correlated versus the alternative hypothesis of error terms that has autocorrelation. The test relies upon the number of observations and the number of parameters. The test statistic ranges from 0 to 4, with a value near 2 indicating no autocorrelation in the residuals. A value approaching 0 indicates positive autocorrelation, while a value approaching 4 indicates negative autocorrelation. Two numbers, depending on observations and parameters, results in *dl* and *du*—critical values are then referenced from the Durbin-Watson test statistics table. From 0 to *dl*, the test statistics falling in this zone will mean a positive serial correlation. From *dl* to *du*, autocorrelation cannot be determined. From *du* to "4-*du*", there is no serial correlation. From "4-*du*" to "4-*dl*", it again means indeterminate for serial correlation. Lastly, from "4-*dl*" to 4, it is then said to be a negative serial correlation.

Based on the figures above, the test statistics for monthly and quarterly columns are near zero, indicating positive autocorrelation. The test statistics for the remaining four columns under daily, indicate non-autocorrelation at first glance as they are nearer to 2 than 0. However, upon closer look, it is not. Column (1) and (2) has *dl*=1.718 and *du*=1.820 at a 5% significance level, the values are below *dl* for both, and hence they are said to be positive and serially correlated. For columns (3) and (4), the *dl*=1.697 and *du*=1.841. As they are within this range, autocorrelation cannot be determined. For columns (5) and (6), *dl*=1.503 and *du*=1.696. The test statistics are near zero, and hence positive autocorrelation. The same can be said for columns (7) and (8), with *dl*=1.336 and *du*=1.72.

The Breusch-Godfrey test with lag=1 is used to check on the findings above. The authors also produced a similar autocorrelation test (Breusch, 1978; Godfrey, 1978). The null hypothesis is that there is no serial correlation, while the alternative hypothesis is that there is a serial correlation. From the results above in Panel B, there is autocorrelation. Column (3) and (4) shows the presence of serial correlation as the *p-value* is less than 0.05, which violates the assumption of no serial correlation.

The implication is that while the estimators may remain consistent and unbiased, they will not be efficient anymore—i.e., not BLUE. Serial correlation is also prevalent in time-series data such as this paper. Consequentially, if it does bring in biasedness as well, the result of it will be overestimation in the R^2 value and *t-statistics* value, they both will be higher than it should be. To remedy, future research can implement transformations of variables via Cochrane-Orcutt Procedure or make the static model into a dynamic model by adding lagged regressors (first differences procedure). The latter will also result in further problems, which will be elaborated on in the limitations of chapter 8.

4.2.4 Assumption 4: Residuals are normally distributed [$u_t \sim N(0, \sigma^2)$]

In this assumption, the skewness-kurtosis test or Jarque-Bera test are the two standard methods used to test for normality. The former will be conducted to measure the asymmetry of the distributions. Panel C of table 2 shows the result, where the null hypothesis of *p-value <0.05* implies data follows a normal distribution and alternative hypothesis of otherwise.

Except for columns (5) and (6) under monthly, probability values were all less than 0.05, implying that residuals do not show normal distribution. Despite this—and on the quality of the estimates—the Gauss-Markov Theorem does not require residuals to follow a normal distribution to produce unbiased and efficient (minimum variance) estimates.

As the OLS residuals are not all asymptotically normally distributed, the Newey-West estimator was used for this paper's time-series regression.

4.3 Factor Loadings' Summary Statistics (Multicollinearity Checks)

In this section, the summary statistics of the factor loadings and checks for multicollinearity will be done. The section will be sub-divided into three sections, based on parts 1, 3 and 4 analysis of figure 1 above. Multicollinearity surfaces when one regressor in the model highly correlates with one or more other regressors ($-1.0 < \rho < -0.5$ or $0.5 > \rho > 1.0$). When there is multicollinearity, Allen (1997) found that it leads to undermining of the statistical significance of the regressor. Consequentially, the standard errors will be larger and the less likely it will be for the estimated coefficient to be statistically significant.

Hence, the three sub-sections below will show tests of multicollinearity via variance inflation factor (VIF) and correlations between regressors.

4.3.1 Part 1's CAPM and multifactor analysis factor loadings (monthly returns)

Table 3 below shows no regressors having a high correlation between themselves. Moreover, the mean VIF is low at 1.40. VIF values that are less than 10 indicate no multicollinearity between the regressors. Going by the individual VIF values for each regressor, one can compute the 1/VIF tolerance value for the degree of collinearity. The tolerance values for each were above 0.1; a value with less than 0.1 shows that it is a linear combination of other regressors. Therefore, there is no presence of multicollinearity, i.e., on the monthly returns against the Fama-French 3 factor models, with indices as proxies for factors loadings.

CAPM market factor has the highest mean and the highest standard deviation. On the other hand, the value factor (HML) has the lowest mean. This can be explained by COVID-19 impact on the financial markets from March 2020, and there were subsequent influences owing to emerging COVID-19 variants, like Delta and Omicron variants. That led to high volatility for the market factor while the value stocks underperformed. Value stocks are well represented in sectors that were hardest hit, such as aviation and hospitality, while growth stocks are well represented in technological sectors that performed well during COVID-19.

Table 3. Summary statistics and checks of multicollinearity (part 1 analysis)

	CAPM	SMB	HML	Mean	Std. Dev.	Min	Max
CAPM	1			0.79	4.15	-15.28	14.70
SMB	.42	1		0.25	2.09	-7.33	5.84
HML	.36	- .15	1	-0.39	3.02	-8.98	10.92
VIF	1.54	1.37	1.30				
Mean VIF	1.40						

Note. p-values in all tables given without zero before decimal point when value less than one. All statistics given in two decimal places; figures given with zero before decimal point when value less than one, but statistics can exceed 1. i.e., all statistics except for proportion, correlation and significance level have zero before decimal point when value less than 1. Reporting of statistics in accordance with APA style.

4.3.2 Part 3's Expanded multifactor analysis factor loadings (daily returns)

In the same vein, table 4 below shows the correlation and VIF values for each of the regressors. For single (CAPM) factor, Fama-French 3 factor and Carhart 4 factor model, the dataset of three-factor loadings from Prof. French's data library was used—hence, panel A shows the values. For Fama-French 5 factor and Fama-French 5+Momentum factor, the five-factor loadings and momentum factor loadings were used—therefore, look at panel B.

Table 4. Summary statistics and checks of multicollinearity (part 3 analysis)

Panel A		RmktRf	SMB	HML	WML			Mean	Std. Dev.	Min	Max
	RmktRf	1						0.02	1.51	-12.01	8.54
	SMB	-.51	1					0.02	0.54	-3.29	1.94
	HML	.40	-.21	1				-0.10	0.70	-3.03	2.49
	WML	-.33	.29	-.82	1			0.09	1.03	-4.32	3.66
	VIF	1.56	1.42	3.31	3.18						
	Mean VIF	2.37									
Panel B		RmktRf	SMB	HML	RMW	CMA	WML	Mean	Std. Dev.	Min	Max
	RmktRf	1						0.06	1.49	-9.62	8.32
	SMB	-.49	1					-0.02	0.61	-5.27	2.00
	HML	.21	.24	1				-0.11	0.75	-3.10	2.39
	RMW	-.02	-.29	-.49	1			0.03	0.26	-1.52	0.96
	CMA	-.08	.20	.74	-.24	1		-0.05	0.33	-1.91	1.17
	WML	-.26	-.08	-.75	.28	-.50	1	0.09	1.01	-4.32	3.66
	VIF	1.86	1.67	5.8	1.48	2.85	2.43				
	Mean VIF	2.68									

In panel A, the value factor negatively correlates with the momentum factor. In panel B, the value factor has a high inverse correlation with the momentum factor and a high correlation with the investment policy factor. Conferring to section 4.3.1, the performance of value stocks was touched upon. This confirms the expectations of high correlational value with momentum and investment policy factors. When those value stock companies are doing poorly, they are the losers but are still conservative in their investment policy. The value factor is calculated via a high Book-to-Market ratio (value) minus stocks with a low Book-to-Market ratio (growth).

To confirm whether there is multicollinearity, the mean VIF of panel A and panel B are below 10 still. The relevant 1/VIF values for each regressor were all above 0.1, which shows no multicollinearity as well.

On factor summary statistics, panel A of table 4 is like table 3 of sub-section 4.3.1. Market factor has the highest mean and volatility; value factor has the lowest mean. Both panels correspond to the study of 180 days before COVID-19 (February 20, 2020) and 180 days after. The period of COVID-19 dramatically depressed the factor loadings (the mean column) you see in both panels. Comparing the results with section 4.3.1, the mean values in those part 1 analyses were higher due to a longer period that stretches from January 2017 to May 2022.

4.3.3 Part 4's Macroeconomic factor analysis factor loadings (quarterly returns)

In this final part on macroeconomic factor loadings, the checks are done on the four factors previously mentioned: balance of payment current account (Euro area), unemployment, inflation (CPI), and short-term interest.

Table 5. Summary statistics and checks of multicollinearity (part 4 analysis)

	BOPSur	UnempSur	CPISur	STIntSur	Mean	Std. Dev.	Min	Max
BOPSur	1				-0.0139	0.7114	-2.2030	1.6570
UnempSur	-.12	1			-0.0030	0.0482	-0.1007	0.1346
CPISur	-.09	-.45	1		-0.0004	0.0475	-0.1209	0.0691
STIntSur	-.05	.09	-.06	1	-0.0001	0.0006	-0.0022	0.0014
VIF	1.05	1.31	1.29	1.01				
Mean VIF	1.16							

Note. Statistics given in four decimal places, in this table, due to near zero figures and need for precision.

Table 5 above shows no regressors having a high correlation. The mean VIF of 1.16 is also low, while the individual 1/VIF for the regressors was all above 0.1. These values prove no presence of multicollinearity within the regressors. Interestingly, the mean of all the regressors was below 0. The factor loadings were all surprise ratios, showing that all four regressors' average surprise ratios were negative. i.e., the expectations of each factor's quarterly rate from Q4 2010 to Q4 2021 were on average lower than the actual rate. Except for the balance of payment factor, the other three factors had close to zero for their maximum surprise ratio in the period studied.

5. Methodology

In this chapter, the model specifications for parts 1 through 4 will be specified in detail. These four parts all stem from the two-pronged approach: portfolio performance differences and performance resilience. The total returns for the regressand, regardless of the return's frequency, are specified by Morningstar to be the following:

$$Total\ Return_t = \frac{\Delta price_{t,t-1} + reinvesting_t + income, gains\ distributions_t}{starting\ price_t} \tag{1}$$

Total return at time t equals to the change in price, plus reinvestment and income, gains distribution over the same period, divided by the price at the beginning of period t.

5.1 Responsible Investing—Portfolio Performance Differences

In this section, the use of the multifactor performance model (CAPM and Fama-French 3 factor) and performance measures will be specified to aid in determining whether there are any performance differences with respect to responsible investing through a clean portfolio.

5.1.1 CAPM and Multifactor Analysis (Monthly Returns)

To start part 1, the single factor model, also known as the capital asset pricing model (CAPM), will be used. CAPM is an equilibrium model that dictates the market exposure as the risk measure to explain risk-adjusted returns performances. The CAPM model has long been found in early research, going back to the 1960s (Treynor, 1961; Sharpe, 1964; Lintner, 1965).

$$R_{it} - R_{ft} = \alpha_i + \beta_1(R_{mt} - R_{ft}) + \varepsilon_{it} \tag{2}$$

Market factor (R_{mt}-R_{ft}) is the systematic risk, while the residual, ε_{it} is the idiosyncratic risk. The beta, β, captures the sensitivity of the market exposure regressor. The regressand on the left-hand side of equation 2 is the excess return of portfolio *i* at time *t*. The constant is Jensen's alpha which will be elaborated in detail in section 5.1.2 of the performance measures. Despite its simplicity, research has proven against CAPM—precisely due to its simplicity—and came up with alternative models, such as arbitrage pricing theory and other multifactor models (Fama & French, 2004).

$$R_{it} - R_{ft} = \alpha_i + \beta_1(R_{mt} - R_{ft}) + \beta_2(SMB_t) + \beta_3(HML_t) + \varepsilon_{it} \tag{3}$$

Fama and French (1993) came up with a 3-factor model (equation 3), by adding two fundamental factors of stocks: size and value. Size factor, SMB_t, represents the outperformances of small cap companies over large cap companies in the long run. Value factor, HML_t, represents value stocks (those with high book-to-market ratio) having outperformance over the growth stocks (those with low book-to-market ratio, low B/M). These two additions to the market factor have spawned a new breed of investors who trade via factor investing styles in this century. The size and value factors are generally undisputed in terms of empirical evidence of its outperformance and seen as economic anomalies to trade on.

Together with CAPM, these two models will be used to regress upon the time-series data of factor loadings via indices as factor proxies, and the monthly returns as the regressand.

5.1.2 Performance Measures (Monthly Returns)

In part 2 analysis, the use of 9 different performance measures is detailed here. Monthly returns from January 2017 to May 2022 (*n* = 65) will be used. The performance measurements differ based on their risk component: volatility, σ; beta, β; Lower Partial Moments (LPM), or the drawdowns. Sharpe ratio is one prominent example of the use of volatility as its risk measure, i.e., the full volatility component. Ratios that use betas, such as the Treynor ratio, use the systematic risk component as its measures. Other more complex form of risk component measures such as LPMs—which considers the downside deviation—and drawdowns, which emphasise the losses incurred over the investment horizon, are also included in this paper to complement the other performance measures.

Sharpe Ratio

Sharpe ratio is a risk-adjusted returns measure used to examine the performances of the portfolio's returns (Sharpe, 1966). However, this performance measure is only good if the returns are normally distributed. In addition, as it examines for the relationship of the returns (risk premiums) and the standard deviation (full volatility component), it leads to the issue of whether the volatility was driven due to upside (which is good) or downside risks. Equation 4 below shows the formula for Sharpe ratio. In the numerator, $\overline{R_p - R_f}$, the average monthly excess returns over 65 months were computed. That figure is then divided by the standard deviation of the excess returns over the said period. Risk-free rate, R_f, used here is the same—1-month Euribor.

$$Sharpe\ Ratio = \frac{\overline{R_p - R_f}}{\sigma_{(R_p - R_f)}}\ ;\ R_f\ being\ 1\ month\ Euribor \tag{4}$$

Treynor Ratio

To address the issue of the Sharpe ratio, the Treynor ratio was used to look specifically at the systematic risk component instead of the full volatility component (Treynor, 1965). Here, the beta, β_{R_p}, of the portfolio returns is used instead of the standard deviation. Beta figure stems from the CAPM in equation 2.

$$Treynor\ Ratio = \frac{\overline{R_p - R_f}}{\beta_{R_p}} \tag{5}$$

This measure informs the investor of the excess return of their portfolio when adjusted for the risk concerning the systematic risk component specifically. Hence, the lower the systematic risk of the portfolio or the asset class is, the better the risk-adjusted performance is.

Information Ratio

Apart from the above two ratios, the information ratio is also one of the most common measures computed by investors. It is like the Sharpe ratio, but instead of using the risk-free rate, the benchmark return is chosen. In this instance, the benchmark is the MSCI Europe IMI index. The average excess return over benchmark is then divided by the tracking error, $\sigma_{(Rp\text{-}Rb)}$. The relevance of this ratio is minimal as it is only helpful for active portfolio management.

$$Information\ Ratio = \frac{\overline{R_p - R_b}}{\sigma_{(R_p - R_b)}}\ ;$$
$$R_b\ being\ return\ of\ benchmark\ MSCI\ Europe\ IMI \tag{6}$$

Sortino Ratio (LPM-based measure)

The classical approach (Sharpe, Treynor ratio) will be insufficient to understand the fat tail risk of the portfolio. Sharpe ratio may lead to instances of underestimating risk and overestimating performances. Instead, using lower partial moments (LPM) for downside deviations is an appropriate risk measure. To obtain LPM, a minimally acceptable return (MAR) must be determined: zero, risk-free rate or the mean return. Here, the risk-free rate is the MAR. Concerning the higher order moments, the LPM uses 0 to 2. LPM with the order of 0 for the shortfall probability, order of 1 for expected shortfall and order of 2 for the semi-variance measure (Eling & Schuhmacher, 2007).

$$Total\ Downside\ Deviation\ (TDD) = \sqrt{\frac{1}{N} * \sum_{i}^{N} \min\left(0, R_p - R_f\right)^2} \tag{7}$$

Equation 7, which illustrates the formula for total downside deviation, is akin to the LPM-based measure. That equation accounts for the negative deviations of each month, with 0 deviations if the excess return is higher than MAR over 65 months (*N*=65). Equation 8 specifies the Sortino ratio used. Sortino and van der Meer (1991) came up with the Sortino ratio, allowing the readers to know what the excess return is, with respect to the total downside deviation (LPM of order 2), i.e., how it performed in the face of its downside risk. The average excess return over the 65 months is divided by the LPM order of 2 (the TDD).

$$Sortino\ Ratio = \frac{\overline{R_p - R_f}}{TDD} \tag{8}$$

M^2 Measure

Despite the popularity of the Sharpe and Treynor ratio, they both have a limitation when it comes to comparing portfolio performances: one can only rank them but not know how much better they are performing against a market portfolio (the benchmark). Again, the MSCI Europe IMI index is the benchmark. To address this limitation, the M^2 or Modigliani-Modigliani measure was created to expand on the Sharpe ratio by adjusting the risk of the portfolio to relative against the risk of a benchmark (Modigliani & Modigliani, 1997). Moreover, M^2 is more intuitive to understand than Sharpe or Treynor ratio as it is in percentages, rather than being without a unit of measure. For instance, two portfolios have Sharpe ratios of 0.70 and -0.70. While one can rank them, it is not intuitive to interpret the Sharpe ratio of -0.70 on how much worse it is compared to the one with 0.70.

$$M^2\ (sharpe\ variant) = \frac{\overline{R_p - R_f}}{\sigma_{(R_p - R_f)}} * \sigma_b + R_f \tag{9}$$

$$M^2\ (treynor\ variant) = \frac{\overline{R_p - R_f}}{\beta_p} * \beta_b + R_f \tag{10}$$

Two different M^2 measures were used. Equation 9 uses the Sharpe ratio variant by accounting for the entire risk component, multiplying with the benchmark volatility. Equation 10 uses the Treynor ratio variant by looking at only the systematic risk component and multiplying it with the systematic risk of the benchmark. Both equations will then be increased with the risk-free rate. It is the M^2 measure, not the M^2 ratio, used. The latter is the market's volatility to the portfolio's volatility.

Drawdown-based Risk Measure

The last alternative for risk measure used is drawdown. A portfolio's drawdown computes the losses incurred over a given period—time under water—before it bounces back to the previous peak. Equation 11 provides the method used to calculate drawdown at time *t*. At each monthly period, the drawdown is the minimum of either 0—if the value is higher than the prior value—or the negative value, which is the current value minus the previous peak value and divided by the previous peak value.

$$Drawdown_t\ (DD_t) = \min\left(0.\frac{p_t - p_{max}}{p_{max}}\right);$$
$$where\ p_t\ is\ current\ value\ of\ portfolio, p_{max}\ is\ the\ historical\ peak \tag{11}$$

$$Average\ DD_t = \frac{1}{T}\sum_{t=1}^{T} DD_t \tag{12}$$

$$Maximum\ DD_t = \max\left(DD_t\right) \tag{13}$$

The average drawdown, equation 12, is the mean drawdown across 65 months. The maximum drawdown, equation 13, is the most significant drawdown to have occurred over the 65 periods. With these drawdowns, one can compute drawdown-based risk measures to understand how the portfolio performed over 65 months by looking at it through the lens of its losses. Three ratios are used here: Sterling ratio, Calmar ratio and Burke ratio.

$$Sterling\ Ratio = \frac{Average\ annualised\ R_p - Average\ annualised\ R_f}{|Average\ annualised\ maximum\ DD_t|} \tag{14}$$

Sterling ratio, a measure more commonly used for hedge funds, was first proposed by Deane Sterling Jones company in 1981. There are several variants of the Sterling ratio. One can use the average or the maximum drawdown to measure risk, and one can use compound returns for its numerator, for instance (Kestner, 1996). This paper uses the average annualised variant of excess return over the average annualised maximum drawdown—see equation 14. As it is annualised, the period studied is January 2017 till December 2021 (*n=60months, or five years*), instead of the previous 65 months. The five months of data for 2022 have been removed as it would complicate the interpretations by trying to annualise five months' returns into one year.

$$Calmar\ Ratio = \frac{Annualised\ \left(R_p - R_f\right) last\ 36\ months}{|maximum\ DD_t\ last\ 36\ months|} \tag{15}$$

The Calmar ratio (equation 15) is another similar drawdown-based measure by slightly modifying the Sterling ratio. The last 36 months are taken from January 2019 to December 2021. These 36 months of excess returns are annualised and divided by the maximum drawdown over the last 36 months (Young, 1991). The critical difference between Calmar and Sterling is the former brings in a 'normalised' interpretation of risk measures for the last 36 months. It smooths the periods of underperformances and overperformances, which is an advantage over the Sterling ratio.

$$Burke\ Ratio = \frac{Average\ annualised\ R_p - Average\ annualised\ R_f}{\sqrt{\sum_{t=1}^{DD} DD_t^2}} \tag{16}$$

Like the Sterling ratio, the Burke ratio attempts to make the Sharpe ratio sharper by changing its denominator. While the numerator is the same as the Sterling ratio in equation 14, the Burke ratio first takes the sum of the squared annualised drawdown of the 60 months (January 2017 to December

2021) and then the square root of that value. The Burke ratio offers a different interpretation than Sterling, where the latter looks at the maximum drawdown to have occurred each year by looking at just the annualised drawdown instead of the maximum drawdown. With an annualised period of 5 years, the first annualised year (*t=1*) becomes the base year, i.e., the annualised DD2t=1 is 0%. The subsequent four years are then computed to see if they will be lower than the initial peak (value at t=1)—resulting in annualised DD2 figure—or it will be the new peak which surpasses the prior peak value—which makes the annualised DD2 for that period to be 0 as it is the new peak instead of a negative deviation.

5.2 Responsible Investing—Resilience over COVID-19 (Event Study)

Moving on from performance differences, this paper will look at the portfolio performance resilience over COVID-19 as an exogenous shock. Section 5.2.1 is an event study using daily returns to study before (180days before February 2020) and during COVID-19 (February 2020 onwards for 180days). Conferring to figure 3 of chapter 3 conceptual framework, five different factor models will be used. Next, section 5.2.2. will detail the model specification of macroeconomic factor analysis.

5.2.1 Single Factor and Multifactor Analysis (Daily Returns)

This sub-section refers to the part 3 approach of expanded multifactor analysis to study performance resilience during COVID-19. Referring to section 5.1.1, the use of CAPM and Fama and French (1993) 3-factor model is also specified here in Equations 17 and 18. A dummy variable, $D_{covid,t}$, has been added to the model specifications for all five models to understand the interaction of COVID-19 with the portfolio's excess returns. The coding of 0 for 180days before February 20, 2020, is in effect and 1 for 180 days from February 20, 2020. The regressand is the total return from Morningstar in daily frequency. The Greek letter alpha, for portfolio *i*, represents Jensen's alpha of the regression. Considering for the systematic risk in the model, a positive alpha indicates outperformance and vice versa.

$$R_{it} - R_{ft} = a_i + l_1(D_{covid,t}) + b_i(R_{mt} - R_{ft}) + \varepsilon_{it} \tag{17}$$

$$R_{it} - R_{ft} = a_i + l_1(D_{covid,t}) + b_i(R_{mt} - R_{ft}) + s_i(SMB_t) + h_i(HML_t) + \varepsilon_{it} \tag{18}$$

$$R_{it} - R_{ft} = a_i + l_1(D_{covid,t}) + b_i(R_{mt} - R_{ft}) + s_i(SMB_t) + h_i(HML_t) + w_i(WML_t) + \varepsilon_{it} \tag{19}$$

Next, equation 19 shows a momentum factor, WML_t, added to the 3-factor model—an anomaly from Carhart (1997). Equation 20 illustrates the expansion of Fama-French 3-factor model into Fama-French 5-factor model (Fama & French, 2015) by adding two quality factors: profitability, RMW_t; and investment policy, CMA_t. The profitability factor was argued by the authors that stocks of companies with high operating profitability perform better than those of low profitability in the long run. The investment policy factor was argued that companies who were conservative in their investment policy (low total asset growth) have above average returns than those who were aggressive in their investment (high total asset growth). Equation 21 is the final multifactor by adding the momentum factor into the Fama-French five-factor model.

$$R_{it} - R_{ft} = a_i + l_1\left(D_{covi\ \ ,t}\right) + b_i\left(R_{mt} - R_{ft}\right) + s_i(SMB_t) + h_i(HML_t) + r_i(RMW_t) + c_i(CMA_t) + \varepsilon_{it} \tag{20}$$

$$R_{it} - R_{ft} = a_i + l_1\left(D_{covid,t}\right) + b_i\left(R_{mt} - R_{ft}\right) + s_i(SMB_t) + h_i(HML_t) + w_i(WML_t) + r_i(RMW_t) + c_i(CMA_t) + \varepsilon_{it} \tag{21}$$

Blitz et al. (2016) highlighted inadequacy in their study of the Fama-French five-factor model. Namely, the momentum factor from Carhart (1997) was excluded yet had been widely researched for over twenty years. This was why the momentum factor had been added as a sixth factor. However, more factors do not necessarily mean better (improving the model's fit to the data). Especially with six different factors, each additional regressor can lead to overfitting of the model, i.e., the new variable improves the model fit, but it does not add significant value.

5.2.2 Macroeconomic Factor Analysis (Quarterly Returns)

This sub-section refers to the final approach toward understanding the resilience of a clean portfolio versus a dirty portfolio during COVID-19. Equation 22 shows the use of the balance of payment current account (BOP) and unemployment rate (UNEMP) as the regressors.

$$R_{it} - R_{ft} = \beta_0 + \beta_1(BOP_t) + \beta_2(UNEMP_t) + \varepsilon_{it} \tag{22}$$

Equation 23 takes it further by adding two macroeconomic factors: consumer price index (CPI) and short-term interest rate (STINT). CPI serves to be a macroeconomic factor for inflation. All four macroeconomic factors are computed for their factor loadings by looking at their respective surprise ratio: (actual rate – expected rate)/expected rate. The factor loadings are taken from OECD data for the Euro area.

$$R_{it} - R_{ft} = \beta_0 + \beta_1(BOP_t) + \beta_2(UNEMP_t) + \beta_3(CPI_t) + \beta_4(STINT_t) + \varepsilon_{it} \tag{23}$$

6. Results

In this chapter, the results from the regression analysis of parts 1, 3 and 4 and performance measures of part 2 will be presented in the following sections below. For the regression analysis, a new regressand will be created by taking the difference between the returns of a clean portfolio (article 8 & 9 ETFs) and a dirty portfolio (article 6 ETFs).

6.1 Part 1 CAPM, Multifactor Analysis (Monthly Returns)

In this section, the results of the part 1 analysis of the performance difference between a clean and a dirty portfolio will be described in detail. The summary statistics of the monthly returns will first be presented before moving on to the results of the regression analysis.

6.1.1 Summary statistics

From this summary statistics of table 6, the average returns throughout the 65 months for the clean portfolio are higher than for the dirty portfolio. The difference between them is positive, confirming higher returns when an investor conducts responsible investing. In addition, the volatility of a clean portfolio is also lower than a dirty portfolio. The dirty portfolio also differs by having a fatter tail risk, as seen from its higher-order moment of skewness. It is slightly more negative than the clean portfolio, suggesting the dirty portfolio has slightly more tail risk exposure—i.e., more likely to experience sudden significant negative returns. On the kurtosis of both portfolios, both are more than 3, which indicates that both (leptokurtic) already exhibit fat tails in their return distribution. A dirty portfolio's higher kurtosis implies that while its returns are more left-skewed than a clean portfolio's, it has more returns occurring closer to its centre than a clean portfolio. The range and interquartile range of a dirty portfolio is slightly more than a clean portfolio.

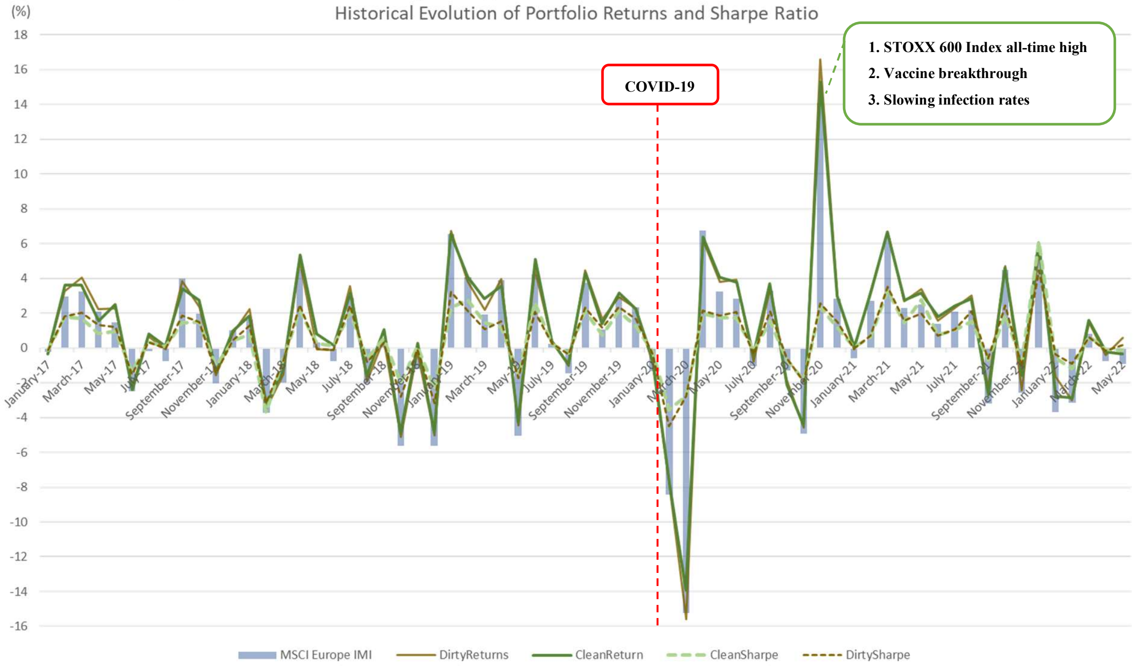
Table 6. Summary statistics of monthly returns, with percentages as unit of measure

Time period from January 2017 to May 2022 (n = 65 months)											
	Min	Q1	Median	Q3	Max	Mean	Std. Dev.	Skew.	Kurt.	IQR	Range
Art. 6	-15.59	-1.01	1.57	3.43	16.60	1.04	4.20	-0.36	7.74	4.43	32.19
Art. 8,9	-13.94	-0.91	1.54	3.41	15.29	1.11	4.02	-0.34	6.59	4.31	29.23
Diff.	-1.32	-0.16	0.13	0.26	1.65	0.07	0.45	-0.13	5.84	0.42	2.96

Note. Article 6 is the equal-weighted dirty portfolio of 51 ETFs, while Article 8 and 9 is the combined equal-weighted clean portfolio of 17 and 3 ETFs, respectively. Skew represents skewness, Kurt represents kurtosis, and IQR for interquartile range. Diff. row stands for the difference between clean and dirty portfolio returns.

Based on the preliminary analysis of just the summary statistics, one can conclude that: 1. the dirty portfolio performs poorer than the clean portfolio, 2. it is more volatile and subjected more to tail risk (extreme returns) than a clean portfolio, and 3. despite its volatility, the dirty portfolio returns seems to be centred closer to its median more often in the past than the clean portfolio.

Figure 6. Monthly Returns and Monthly Sharpe Ratio Movements



Note. Monthly portfolio excess returns and MSCI Europe IMI Index returns given in percentages as unit of measure. Monthly Sharpe ratios computed based on cross-sectional ETFs' returns. Red vertical line shows start of COVID-19 market impact; consensus puts it on 20 February 2020. Not all months displayed on horizontal axis to avoid overcrowding the graph.

MSCI Europe IMI Index was chosen as it represents small to large-cap constituents. Figure 6 depicts the portfolios' monthly excess returns above the risk-free rate and monthly Sharpe ratio movements. The Europe market, represented by the blue bars, illustrates the index returns movement as a proxy for the European financial market. On certain months, the index returns were near zero, which explains why some blue bars cannot be seen in the figure.

The drift and diffusion are largely similar between both portfolios. Past the red vertical dashed line, the portfolios' return and Sharpe ratio slide down dramatically due to the COVID-19 stock market crash globally. Markets termed it the Black Monday of 2020 as stock markets continued to plunge on 9 and 16 March. The bear market was short-lived as, by the end of April 2020, stock exchanges globally re-entered the bull market. In the graph, May 2020 shows the portfolios and the Europe market re-entering the bull market.

In November 2020, figure 6 showed portfolios return and the Europe market reaching its highest peak. This could be due to the following factors: the pan-European STOXX 600 reaching its all-time high, a surge in Eurozone stocks due to vaccine breakthrough news and slowing infection rates from COVID-19 in Europe, allowing governments to ease restrictions.

6.1.2 Regression results

In this sub-section, the results of time-series regression for the 65 months are detailed in table 7. The CAPM single factor model was regressed first before adding the size (SMB) and value (HML) factor into the model. In the CAPM model, the clean portfolio has a Jensen's alpha—higher abnormal return—of 0.3488 versus the dirty portfolio's 0.2497. Both alphas are positive and statistically significant at 99% confidence, i.e., they both exhibit outperformance over the MSCI Europe IMI market as the benchmark. The difference between the returns of the clean and dirty portfolio also confirms the better performance of the clean portfolio. The difference's alpha is positive and statistically significant at 90% confidence. Moreover, the clean portfolio is marginally less sensitive to the European market than the dirty portfolio—both coefficients for the market factor (column 2) are statistically significant as well.

Table 7. Regression results of monthly returns, with percentages as unit of measure

	Alpha (1)	R_m-R_f (2)	SMB (3)	HML (4)	Adj. R^2 (5)
CAPM					
Art. 6	0.25 (0.07)***	1.00 (0.03)***			0.98
Art. 8 & 9	0.35 (0.07)***	0.96 (0.02)***			0.98
Difference	0.10 (0.06)*	-0.04 (0.02)**			0.15
FF3					
Art. 6	0.31 (0.06)***	0.99 (0.02)***	-0.04 (0.03)	0.09 (0.02)***	0.99
Art. 8 & 9	0.35 (0.06)***	0.98 (0.02)***	-0.09 (0.03)***	-0.01 (0.02)	0.98
Difference	0.04 (0.04)	-0.01 (0.01)	-0.06 (0.02)**	-0.10 (0.02)***	0.51

Note. Article 6 is the equal-weighted dirty portfolio of 51 ETFs, while Article 8 and 9 is the combined equal-weighted clean portfolio of 17 and 3 ETFs, respectively. CAPM stands for the single factor model. FF3 stands for Fama-French 3 factor model. Adj. R^2 represents the adjusted R-squared value, i.e., the goodness of fit of the model to the data. *, **, *** indicates *p-values* significance at 10%, 5% and 1% respectively. Figures in parentheses are Newey-West standard errors and are displayed beside its corresponding test statistics.

In the Fama-French 3-factor model, a clean portfolio again performed better. Despite adding two more regressors, the alpha is positive and significant at 99% confidence (0.35 versus 0.3075). Both portfolios also have significance, with the market factor at nearly 1.0. On size factor, a dirty portfolio has no statistical significance while a clean portfolio has negative relevance. The dirty portfolio exhibits positive significance for the value factor, while the clean portfolio has no statistical significance. The difference in returns supports the above results partially. The size factor has negative significance at 95% confidence, while the value factor seems to show negative significance at 99% confidence. There is no statistical significance in the difference in returns for alpha.

Based on the regression analysis of the two models, one can confirm that the clean portfolio does perform better than the dirty portfolio. While it is not affected by value stocks' outperformance, the clean portfolio is marginally affected negatively when small-cap stocks have outperformance. The coefficient estimates for difference show clean portfolio exposure towards small cap firms and growth stocks. Both underperformances of small-cap firms and growth stocks led to negative and significant coefficient estimates. More on this in chapter 7 discussion.

6.2 Part 2 Performance Measures (Monthly Returns)

Twelve different performance measures were computed to understand how the clean portfolio performed in risk-adjusted returns but with varying risk measures. The four different risk measures were previously expanded upon in sub-section 5.1.2. The four risk measures that were used to build these twelve performance measures are:

- Volatility
- Beta (the systematic risk)
- Lower partial moments (the downside deviations)
- Drawdowns (the losses incurred over the horizon)

From table 8, most of the performance measures (9 out of 12) showed clean portfolio performs better, i.e., 75% of the time, the performance measure says the clean portfolio outperforms the dirty portfolio. The Treynor ratio, M2 measure with Treynor ratio variant, and Burke ratio, on the other hand, showed clean portfolio underperformed the dirty portfolio instead.

Table 8. Performance measures based on monthly returns between January 2017 to May 2022 (n = 65 months)

Panel A	Sharpe Ratio	Treynor Ratio	Jensen's Alpha	Information Ratio	Sortino Ratio
Art. 6	0.27	1.16	0.25***	0.37	0.44
Art. 8,9	0.28	1.16	0.35***	0.51	0.51

Panel B	M^2 (Sharpe)	M^2 (Treynor)	Avg. DD. (monthly)	Max. DD. (monthly)	Sterling Ratio	Calmar Ratio	Burke Ratio
Art. 6	23.90	0.73	-0.06	-0.24	1.38	0.82	1.27
Art. 8,9	24.11	0.72	-0.05	-0.22	1.69	0.94	0.54

Note. Article 6 is the equal-weighted dirty portfolio of 51 ETFs, while Article 8 and 9 is the combined equal-weighted clean portfolio of 17 and 3 ETFs, respectively. Jensen's alpha, ***, indicate *p*-values significant at 1%. Drawdown-based (DD) measure computed with period January 2017 to December 2021(n=60). Avg. DD. stands for average drawdown; Max. DD. stands for maximum drawdown. Both drawdowns and M^2 measure are reported in percentages as unit of measure. Jensen's alpha is taken from CAPM constant of the single factor model of part 1 analysis.

First, this paper will look at the classical performance measures. Sharpe ratio above 1.0 means the portfolio offers excess returns after considering the full volatility component. The clean and dirty portfolios have a Sharpe ratio below 1.0, and the clean portfolio has a higher ratio of 0.2756. However, looking at the Sharpe ratio is insufficient as it assumes a normal distribution, which the returns so far have found not to be normally distributed—both portfolio returns have fat tails. Treynor ratio, which accounts for the systematic risk component only, showed that both portfolios performed better when using the beta of the MSCI Europe IMI index. While both are above 1.0, the dirty portfolio has a higher ratio of 1.1642, i.e., the dirty portfolio offers higher returns—after accounting for the risk-free rate—of 0.096.

Jensen's alpha stems from the single factor model of sub-section 6.1.2, while the information ratio is calculated based upon the same benchmark's returns. The latter allows you to peg the portfolio returns relative to the benchmark and see how much the portfolio exceeded the benchmark. The information ratio of a clean portfolio is higher at 0.5059. This suggests that when the European market (proxied via the MSCI Europe IMI index) moves by 1%, the clean portfolio of Article 8 and 9 ETFs provides further excess returns than investing in the dirty portfolio.

To account for the limitations of the Sharpe and Treynor ratio, see sub-section 5.1.2, the M^2 measures were created to address them. With the Sharpe variant of M^2 measures, a clean portfolio had 24.1% versus a dirty portfolio of 23.9%. This is logical as the Sharpe ratio of the clean portfolio was higher than the dirty portfolio, and the volatility of benchmark and risk-free rate is the same in the M^2 measure for clean and dirty portfolios. M^2 measure has percentages as its unit of measure, allowing the investor to rank their portfolio and compare their performances. As such, when accounting for the benchmark volatility, the clean portfolio had a 0.21% higher risk-adjusted return than the dirty portfolio. Moving onto the M^2 measure with the Treynor variant, the clean portfolio had a lower measure than the dirty portfolio (0.7169% versus 0.73%). This is because the Treynor ratio of the clean portfolio is more down than the dirty portfolio.

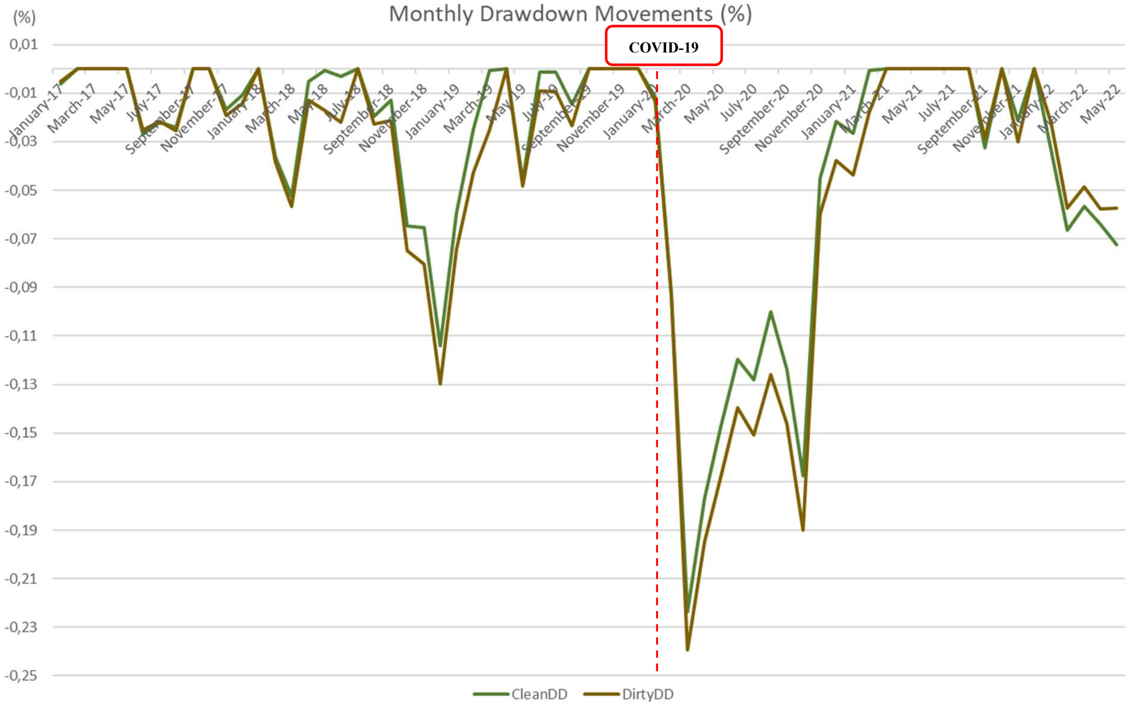
Moreover, the benchmark's beta sensitivity is essentially equal to 1 as it is the market beta. Consequentially, the M^2 measure of the Treynor variant is much lower than the M^2 of the Sharpe variant. Because both portfolios are well diversified (51 ETFs for dirty and 20 ETFs for clean portfolios), the Treynor M^2 is a deserved inclusion. Had it not been well-diversified, there exists a risk of underestimation of the portfolio's riskiness due to idiosyncratic risks which were not part of the risk component.

Sortino ratio, an LPM-based measure, is also included to look at the portfolio's downside risk. Upside risk is not included as the positive deviation is a good risk. Removing for the upside returns movement, the clean portfolio performed better than the dirty portfolio at 0.51 versus 0.44. While one portfolio may have a higher mean return than the other, it does not necessarily mean it is efficient—a higher return could be driven by higher risk. The Sortino ratio confirms clean portfolio is more efficient in driving those returns than a dirty portfolio—given their respective downside deviations.

Lastly, the drawdown-based measure of Sterling ratio, Calmar ratio and Burke ratio were used. The study period between January 2017 and December 2021 is due to the requirements of annualizing monthly returns. Hence n = 60 months instead of 65 months. Cumulating all the drawdowns over the past 60 months, the average monthly drawdown was -5.37% and -6.23%, respectively, for clean and dirty portfolios. Figure 7 illustrates the monthly drawdowns movement of both portfolios. Past the red vertical dashed line, one can see the losses incurred over the horizon dramatically slide down for both portfolios. By the third quarter of 2021, both portfolios had recovered, but the clean portfolio rebounded faster than the dirty portfolio. The maximum drawdown in those periods—the highest monthly loss—

was -22% and -24% for clean and dirty portfolios. These numbers meant that in the past 65 months when the portfolios incur losses, the clean portfolio fared ‘better’ on average—less negative in monthly drawdown. When looking at the maximum loss in March 2020 for both portfolios, clean portfolios suffered lower losses than dirty portfolios. From figure 7, one can see the clean portfolio rebounded better each time following a loss.

Figure 7. Monthly Drawdown Movements



Note. Red vertical line marks start of COVID-19 market impact globally. Consensus puts the start of market slide from 20 February 2020. Not all months are displayed on horizontal axis to avoid overcrowding the graph.

The Sterling ratio looks at the average annualised excess return over the average annualised maximum drawdown. The denominator is in absolute value to prevent division over negative numbers. Because the clean portfolio fared better during losses, its Sterling ratio was higher at 1.7 versus 1.4 for the dirty portfolio. Calmar ratio, which examines the last 36 months, is helpful since COVID-19 occurred two years ago. Calmar ratio smooths the volatility of the past 36months by annualising all 36 months together. In this ratio, a clean portfolio does better than a dirty portfolio at 0.94 versus 0.82. In the Burke ratio, a clean portfolio was computed to have a much lower ratio than a dirty portfolio at 0.54 vs 1.27 for a dirty portfolio. As such, a deeper investigation into why it is so was conducted.

From the summary statistics of monthly returns, a clean portfolio had higher mean returns than a dirty portfolio. To calculate the Burke ratio, the annualised excess returns were taken (see equation 16). That means five different annualised excess returns are used, with the first 12 months (January 2017 to December 2017) being the first year of annualised excess returns. The first year is the base year for calculating the subsequent annualised drawdowns for the 2nd to 5th year. Now, because the mean return for the clean portfolio was higher, its base year value is higher than the dirty portfolio (the peak). In the second year, both portfolios' annualised returns were negative—hence they are losses instead—but they were about the same level. The loss is the trough. When one thinks of a wave, this is the peak-to-trough ratio between the first and second years. The peak-to-trough for the clean portfolio was more significant than the dirty portfolio, which resulted in a larger annualised drawdown and incidentally gave the wrong interpretation at first glance. This phenomenon occurred again in the fourth year (2020 COVID-19). Consequentially, the Burke ratio for the clean portfolio became lesser than the dirty portfolio as its peak-to-trough was larger.

6.3 Part 3 Expanded Multifactor Analysis (Daily Returns)

Like section 6.1, the results of the part 3 analysis of performance resilience between clean and dirty portfolios will be described in detail. The summary statistics of the 360 daily returns will first be presented before moving on to the regression analysis results.

6.3.1 Summary statistics

Table 9. Summary statistics of daily returns, with percentages as unit of measure

	Before COVID-19 market crash (5 June 2019 to 19 February 2020)										
Panel A	Min	Q1	Median	Q3	Max	Mean	Std. Dev.	Skew	Kurt.	IQR	Range
Art. 6	-2.60	-0.28	0.09	0.48	2.14	0.08	0.70	-0.68	5.24	0.75	4.74
Art. 8 & 9	-2.65	-0.26	0.07	0.47	2.30	0.08	0.72	-0.65	5.34	0.73	4.95
Difference	-0.18	-0.04	-0.00	0.04	0.17	0.00	0.06	0.13	3.19	0.08	0.35

	During COVID-19 market crash (20 February 2020 to 5 November 2020)										
Panel B	Min	Q1	Median	Q3	Max	Mean	Std. Dev.	Skew	Kurt.	IQR	Range
Art. 6	-11.91	-0.75	0.07	1.03	8.41	-0.08	2.03	-1.21	10.82	1.78	20.32
Art. 8 & 9	-11.74	-0.79	0.11	1.01	8.73	-0.07	2.06	-1.11	10.44	1.80	20.47
Difference	-0.74	-0.06	0.02	0.08	0.68	0.01	0.16	-0.15	7.54	0.14	1.42

	Entire sample period (5 June 2019 to 5 November 2020)										
Panel C	Min	Q1	Median	Q3	Max	Mean	Std. Dev.	Skew	Kurt.	IQR	Range
Art. 6	-11.91	-0.45	0.08	0.67	8.41	0.00	1.51	-1.59	17.62	1.12	20.32
Art. 8 & 9	-11.74	-0.46	0.08	0.69	8.73	0.01	1.54	-1.46	16.88	1.15	20.47
Difference	-0.74	-0.05	0.00	0.06	0.68	0.01	0.12	-0.07	11.21	0.11	1.42

Note. Article 6 is the equal-weighted dirty portfolio of 57 ETFs, while Article 8 and 9 is the combined equal-weighted clean portfolio of 25 and 4 ETFs, respectively. Skew represents skewness, Kurt represents kurtosis, and IQR for interquartile range. Panel A represents summary statistics of 180 daily returns prior to COVID-19, panel B is 180 daily returns during COVID-19. Panel C represents the summary statistics of entire 360 daily returns of sample.

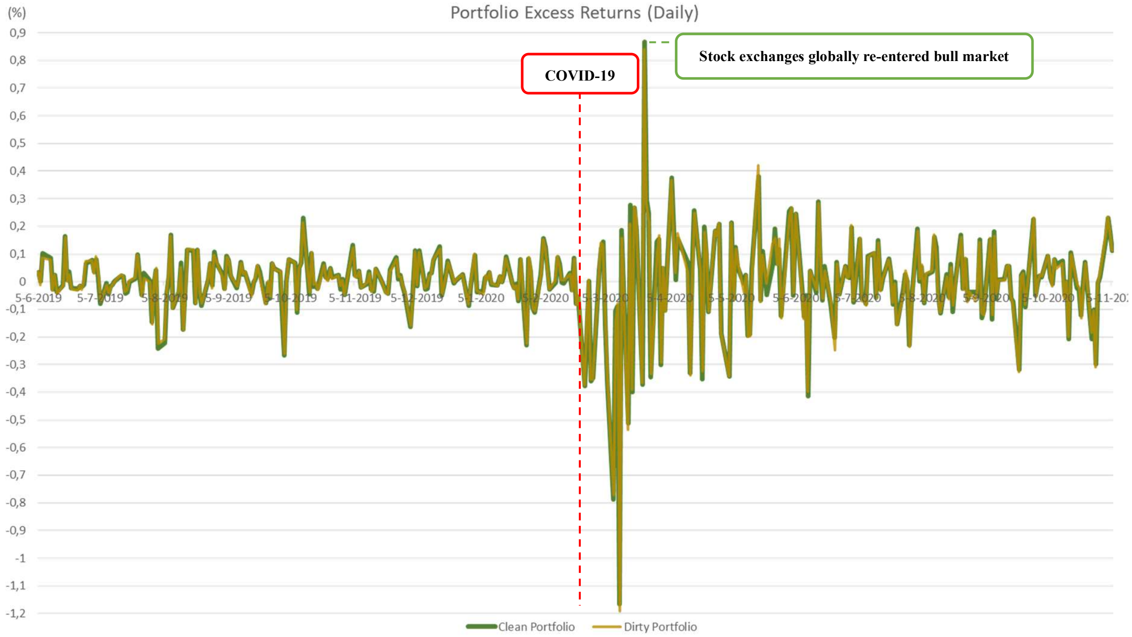
In table 9, the summary statistics are provided via three panels. Panel A for the 180 daily returns before COVID-19, panel B for the 180 daily returns during COVID-19, and panel C for the 360 daily returns of the entire period. Like sub-section 6.1.1, a regressand is added by taking the difference between the returns of clean and dirty portfolios. This is to enable a regression analysis into the performance return differences for its significance.

On average, the clean portfolio has a higher return than the dirty portfolio throughout the entire period and before COVID-19 (0.0808 vs 0.0796 of dirty portfolio mean). During COVID-19, the clean portfolio also fared better by having lower average daily losses than the dirty portfolio. The positive figure in the difference row for all three panels further supports the abovementioned finding. The volatility of the clean portfolio is higher than the dirty portfolio in all three panels this time. From subsection 6.1.1, when the monthly returns (with a period stretching back to January 2017) were examined, the clean portfolio had lower volatility than the dirty portfolio. This shows that during the 360days period studied, the clean portfolio has a higher return and slightly higher volatility. The higher volatility of the clean portfolio explains why it has higher interquartile range and range than the dirty portfolio.

When looking at the higher order moments of 3 (skewness) and higher order moment of 4 (kurtosis), both portfolios' return exhibits tail risk exposure (negative skewness) and fat tails (kurtosis beyond 3). In all three panels, the clean portfolio has slightly less negative skewness than the dirty portfolio. This means a clean portfolio is less likely than a dirty portfolio to experience sudden significant negative returns (tail risk). In addition, a clean portfolio has slightly lower kurtosis than a dirty portfolio. As both portfolios exhibit leptokurtic in their return distribution, they exhibit high 'tailedness' relative to a normal distribution. i.e., they have more returns concentrated around the centre of the distribution. In this instance, both portfolios, due to negative skewness and high kurtosis, will have occasional extreme returns that are at least three or more standard deviations away from the mean, and it will occur on the left side of the curve.

Figure 8 illustrates the excess daily returns (Rp-Rf) for both portfolios. Past the red vertical dashed lines, COVID-19 impact occurs. One can notice that between the first and second quarters of 2020, the daily returns movement of both portfolios exhibits more significant fluctuations. From April 2020, stock markets re-entered the bull market globally, as seen in figure 8. Like figure 6, which demonstrates the portfolios' monthly returns, figure 8 exhibits the portfolio's daily returns and is aligned in illustrating the short-lived COVID-19 market crash.

Figure 8. Daily Returns Movements, in percentages



Note. Red vertical line represents start of COVID-19, 20 February 2020. Vertical axis represents the daily returns movements in percentages as unit of measure. Clean portfolio (green) underlaps dirty portfolio (brown) in graph to show its performance difference.

6.3.2 Regression results

Here, the results of time-series regression from table 10 are detailed. To reiterate, the regressions of CAPM, Fama-French 3-factor model (FF3), and Carhart 4-factor model have regressed with factor loadings of Prof. French's 3-factor loadings and momentum factor loadings dataset. In contrast, the Fama-French 5 factorml (FF5) and FF5+Momentum factor model were regressed using Prof. French's 5-factor and momentum factor loadings dataset. These datasets targeted Europe as a region, with US T-bill as risk-free. Hence, the regressand of daily returns was converted from Euro to US dollar, based on the prevailing day's closing spot rate. The analysis starts first with the market factor (Rm-Rf) and slowly adds size (SMB), value (HML), momentum (WML), profitability (RMW) and investment policy (CMA) factors as additional regressors. A dummy variable is specified in the model to study the interaction between COVID-19 and the regressand.

First, in the single factor model, CAPM alpha for both clean and dirty portfolios showed no significance in out- or underperformance. The clean portfolio has lesser negative and significant estimates during COVID-19 at 90% confidence. i.e., the clean portfolio was more resilient than the dirty portfolio as its estimates were higher by approximately 0.01. On the market factor, the clean portfolio showed no significance in exposure, while the dirty portfolio does at 99% confidence. Although the dirty portfolio has a more negative and significant coefficient estimate for COVID-19, it has a positive and considerable exposure to the market factor. When the market moves by 1%, the dirty portfolio's returns move in the same direction by 0.96. There was no significance in the results when it comes to regressing the CAPM model, with the difference in returns as regressand.

Second, to increase the goodness of fit, the FF3 model is employed. This time, both clean and dirty portfolios show positive and significant outperformance at 90% confidence. The clean portfolio has a marginally higher outperformance than the dirty portfolio, with 3.77% versus 3.74% in for dirty portfolio. Moreover, the clean portfolio was also affected less negatively by COVID-19 than the dirty portfolio. The former has a negative and significant estimate of 0.075 at 90% confidence, while the latter has a negative and significant estimate of 0.083 at 95% confidence. Similar to CAPM model, both portfolios have significant positive exposure towards the market at approximately 0.87, with 99% confidence.

Size factor exposures negatively and significantly affect both portfolios' returns at 99% confidence. This implies the influence of small-cap stocks' outperformance on the ETFs, be it article 6, 8 or 9. i.e., a 1% change of small-cap stocks over large-cap stocks in the indices led to an approximate 0.4 points change in both portfolios returns, ceteris paribus. The difference in returns regressand concurs with the finding, at 99% confidence. When it comes to the value factor, it has a positive and significant influence on the dirty portfolio at 95% confidence. An outperformance of value stocks (high book-to-market ratio) by 1% translates to a 0.095-point increase effect on the dirty portfolio's returns. While there is no significance of value factor loading on the clean portfolio, there is a negative and significant effect in the difference in returns regressand. Further evidencing that a dirty portfolio has exposure towards value stocks, while a clean portfolio has more exposure to growth stocks.

Third, the Carhart 4-factor model adds the momentum factor—12 months winner minus 12 months loser stocks. The Carhart alpha shows positive and significant outperformance of both portfolios, with the clean portfolio having marginally higher outperformance of 2bps (3.73% vs 3.71% dirty portfolio alpha). Like the FF3 model, both portfolios have negative, significant underperformance during COVID-19. The clean portfolio is, again, slightly more resilient than the dirty portfolio as it has a lesser negative coefficient estimate of -0.0733 at 90% confidence. The results of the market factor and size factor are like FF3's model. In the Carhart 4-factor model, adding the momentum factor as an extra regressor was not useful. There was no significance in the coefficient estimate of the momentum factor, while the value factor's estimates were not significant as well.

Fourth, the FF5 model is used. Instead of the momentum factor, two quality factors were added: profitability and investment policy. Adding extra regressors led to the lesser fitting of the model to the data. The adjusted R^2 value for the FF5 and FF5+Momentum factor model was lower than the FF3 and Carhart models. This explains why the FF5 alpha showed no significance in either of the portfolios' regressand. There was also no significance for the COVID-19 dummy variable, value factor, profitability factor and investment policy factor. Size factor showed positive and significant estimates for both portfolios. This is in the opposite direction as size factor loading have, instead, *(continued)*

Table 10. Regression results of daily returns, with percentages as unit of measure

	Alpha	COVID-19	R_m-R_f	SMB	HML	WML	RMW	CMA	Adj. R^2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
CAPM									
Art. 6	0.0280 (0.0228)	-0.0999 (0.0458)**	0.9577 (0.0213)***						0.92
Art. 8 & 9	0.0284 (0.0235)	-0.0897 (0.0472)*	0.9617 (0.0219)						0.91
Difference	0.0005 (0.0047)	0.0102 (0.0128)	0.0041 (0.0070)						0.00
FF3									
Art. 6	0.0374 (0.0200)*	-0.0832 (0.0414)**	0.8740 (0.0256)***	-0.3602 (0.0644)***	0.0952 (0.0417)**				0.93
Art. 8 & 9	0.0377 (0.0200)*	-0.0754 (0.0413)*	0.8694 (0.0240)***	-0.4375 (0.0668)***	0.0659 (0.0416)				0.93
Difference	0.0003 (0.0047)	0.0078 (0.0120)	-0.0046 (0.0080)	-0.0773 (0.0219)***	-0.0293 (0.0124)**				0.11
Carhart									
Art. 6	0.0371 (0.0201)*	-0.0815 (0.0411)**	0.8733 (0.0246)***	-0.3639 (0.0627)***	0.1112 (0.0752)	0.0134 (0.0465)			0.93
Art. 8 & 9	0.0373 (0.0200)*	-0.0733 (0.0409)*	0.8686 (0.0230)***	-0.4420 (0.0639)***	0.0852 (0.0785)	0.0163 (0.0477)			0.93
Difference	0.0002 (0.0047)	0.0082 (0.0116)	-0.0048 (0.0082)	-0.0781 (0.0218)***	-0.0259 (0.0230)	0.0029 (0.0150)			0.11
FF5									
Art. 6	0.0399 (0.0340)	-0.1264 (0.0810)	0.9116 (0.0761)***	0.3744 (0.1243)***	0.1652 (0.1444)		-0.1884 (0.2103)	0.0235 (0.3109)	0.75
Art. 8 & 9	0.0370 (0.0341)	-0.1153 (0.0797)	0.9248 (0.0716)***	0.3235 (0.1187)***	0.1219 (0.1358)		-0.1854 (0.2098)	0.1044 (0.2942)	0.76
Difference	-0.0025 (0.0047)	0.0112 (0.0117)	0.0132 (0.0096)	-0.0509 (0.0152)***	-0.0433 (0.0259)*		0.0031 (0.0358)	0.0809 (0.0481)*	0.14
FF5 + MOM									
Art. 6	0.0439 (0.0347)	-0.1364 (0.0799)*	0.9068 (0.0787)***	0.3846 (0.1216)**	-0.0006 (0.1523)	-0.1367 (0.0805)*	-0.2469 (0.2106)	0.0735 (0.3015)	0.75
Art. 8 & 9	0.0414 (0.0348)	-0.1255 (0.0787)	0.9200 (0.0740) ***	0.3338 (0.1116) ***	-0.0451 (0.1463)	-0.1377 (0.0788)*	-0.2442 (0.2099)	0.1548 (0.2851)	0.76
Difference	-0.0025 (0.0047)	0.0110 (0.0114)	0.0132 (0.0097)	-0.0508 (0.0151)***	-0.0445 (0.0318)	-0.0011 (0.012)	0.0026 (0.0361)	0.0813 (0.0495)	0.14

Note. Statistics given in four decimal places, figures near zero and need for precision. Carhart represents the 4-factor model, FF5 represents the Fama-French 5-factor model. MOM represents the momentum factor. Adj. R^2 represents the adjusted R-squared value, i.e., the goodness of fit of the model to the data. *, **, *** indicates *p-values* significance at 10%, 5% and 1% respectively. Figures in parentheses are Newey-West standard errors and are displayed beside its corresponding test statistics.

(continued) negative estimates in FF3 and Carhart model. It is initially puzzling, but for FF5 and FF5+Momentum factor model, Prof. French has a different dataset for these factor loadings. By adding two quality factors into the regressors, the size factor shows that when small-cap companies exhibit outperformance, the ETFs –regardless of their SFDR article classification—move in the same direction regarding returns. Lastly, looking at the difference between the clean portfolio and dirty portfolio returns as a regressand, there is a negative and significant estimate for the value factor while there is a positive and significant estimate for the investment policy factor—both at 90% confidence.

They are explanatory for clean portfolio returns above dirty portfolio. A clean portfolio has more exposure to companies in sectors with higher sustainability ratings, most prominently the technological (tech) sector. Hence, these articles 8 and 9 ETFs are more exposed to tech stocks, which are growth stocks, when looking at their fundamentals—they have low book-to-market ratios. Hence, the value factor is negative in the direction of clean portfolio returns regressand. i.e., when the value stocks have 1% outperformance over growth stocks, the clean portfolio moves in the opposite direction by 0.04 for the difference in returns. Horstmeyer et al. (2022) found the investment policy factor peculiar as it should mirror the value factor in drift movement. The conservative investment policy is well represented in value companies. But since 2014, CMA factor loading has switched in direction; the CMA cumulative returns have been downward since 2014. This explains why the CMA factor estimate in the 6-factor model, is in the opposite direction of the value factor.

Fifth, the FF5+Momentum factor model combines the momentum factor loading dataset with the 5-factor loading dataset. Like the FF5 model, adding several more regressors does not seem to add value as the model has lower goodness of fit now compared to FF3 and Carhart. The interpretation of the market and size factor is like FF5's in terms of significance and the direction of the estimates. Like FF5, the clean portfolio has a slightly higher positive and significant estimate of 0.92 for market factor, and it is about the same as the dirty portfolio in terms of size factor influence. One thing that is striking is the momentum factor tilts. Both portfolios' regression analyses showed negative and significant estimates of around -0.14, at 90% confidence. This shows the past 12 months of winners became losers during the COVID-19 and had no outperformance. A 1% increase in momentum factor led to a 0.14 decrease effect on both portfolio returns. The reversal was due to COVID-19, owing to the 360days of the period surrounding 20 February 2020—this is not the same as the reversal momentum factor.

6.4 Part 4 Macroeconomic Factor Analysis (Quarterly Returns)

Here in part 4, the quarterly returns of both portfolios were used as the regressand. The period stretched the furthest back to the fourth quarter (Q4) 2010 from Q4 2021 when compared to monthly returns data, which was until January 2017.

6.4.1 Summary statistics

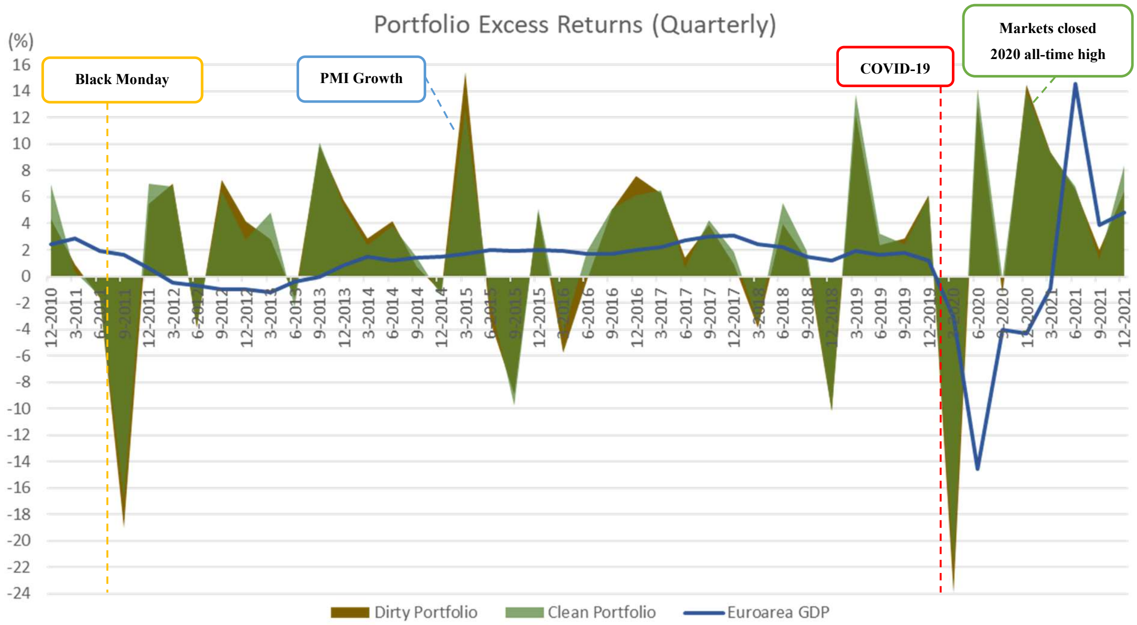
Table 11. Summary statistics of quarterly returns, with percentages as unit of measure

Time period from Q4 2010 to Q4 2021 (n = 45 quarters)											
SFDR	Min	Q1	Median	Q3	Max	Mean	Std. Dev.	Skewness	Kurtosis	IQR	Range
Art. 6	-23.97	-1.18	2.87	6.31	15.41	2.14	7.54	-1.26	5.74	7.49	39.38
Art. 8 & 9	-21.79	-0.16	3.24	6.51	14.20	2.49	7.23	-1.17	5.22	6.67	35.99
Difference	-2.76	-0.45	0.25	1.34	2.63	0.35	1.14	-0.04	2.83	1.79	5.38

Note. Article 6 is the equal-weighted dirty portfolio of 35 ETFs, while Article 8 and 9 is the combined equal-weighted clean portfolio of 10 and 2 ETFs, respectively. Skew for skewness, Kurt for kurtosis, and IQR for interquartile range.

Table 11 shows the summary statistics of quarterly returns from clean and dirty portfolios. The average quarterly return of the clean portfolio concurs with the monthly and daily return being higher than the dirty portfolio. Like section 6.1.1., the volatility of a dirty portfolio is lower than those of a clean portfolio. Here, both portfolio returns distribution is negatively skewed and have higher kurtosis (kurtosis > 3). Hence, the interpretation of the summary statistics for quarterly returns mirrors those of the summary statistics for monthly returns.

Figure 9. Quarterly Returns Movements, in percentages



Note. Red line represents start of COVID-19, 20 February 2020. Vertical axis represents the movements in percentages as unit of measure. Clean portfolio (green) overlaps dirty portfolio (brown) in graph to show its performance difference.

Figure 9 depicts the quarterly excess return movements of both portfolios. The blue curve in the graph illustrates the quarter-to-quarter changes in Euro area GDP, giving one a sense of the general economic development occurring in Europe. Throughout the 45 quarters, the clean portfolio outperformed the dirty portfolio slightly, as evidenced by the alpha from table 12. In the third quarter of 2011, both portfolios' quarterly returns plunged. This shows that the portfolios, although Europe ETFs, are impacted negatively from Black Monday of 2011. During that event, rating agencies downgraded US sovereign debt's credit rating, causing repercussions on US and global financial markets. US indices, such as the NASDAQ, S&P 500, and DJIA[8], plunged by over -5%.

The first quarter of 2015 showed both portfolios reaching new peaks. One possible reason could be the strong PMI[9] growth for Germany and the Eurozone private sector, expanding quickly in February and March 2015. However, after the first quarter of 2015, both portfolios plunged again due to Q2 2015 to Q2 2016 global stock market selloff, driven by several factors such as the Greek debt default and UK's EU membership referendum. In the third quarter of 2020, past the red vertical dashed line, the COVID-19 impact occurred and caused both portfolios and Euro area GDP to slide dramatically. From this point on, the line movements in figure 9 are aligned with figures 6 and 8, which showed a short-lived bear market and rebound from April 2020.

6.4.2 Regression results

Table 12 shows the result of the macroeconomic factor model. The analysis first started by using all four macroeconomic factors' surprise ratios for the regressors. There is no dummy variable for COVID-19. The purpose of the macroeconomic factor model is to examine whether the deterioration of macroeconomic indicators due to COVID-19 influenced the European financial market. From panel B of the table, the clean portfolio has positive and significant outperformance, at 95% confidence. The outperformance of the clean portfolio is higher than those of the dirty portfolio—2.2 versus 1.9 alpha. The difference in portfolio returns between clean and dirty portfolio returns as regressand confirms the finding with positive and significant alpha at 90% confidence.

Table 12. Regression results of quarterly returns, with percentages as unit of measure

	Alpha	BOP	Unemployment	CPI	Short-term Interest	Adj. R^2
Panel A	(1)	(2)	(3)	(4)	(5)	(6)
Art. 6	1.96 (1.11)*	2.26 (1.58)	51.01 (23.79)**			0.09
Art. 8 & 9	2.31 (1.06)**	2.22 (1.53)	49.33 (22.99)**			0.10
Difference	0.35 (0.17)**	-0.04 (0.20)	-1.70 (3.42)			-0.04
Panel B						
Art. 6	1.90 (1.11)*	2.45 (1.64)	59.01 (24.81)**	18.10 (30.08)	391.32 (1260.56)	0.06
Art. 8 & 9	2.21 (1.05)**	2.47 (1.61)	59.56 (23.78)**	23.46 (28.57)	637.15 (1200.26)	0.07
Difference	0.32 (0.17)*	0.02 (0.20)	0.55 (4.51)	5.37 (4.59)	245.01 (337.17)	-0.04

Note. BOP is the balance of payment current account factor. Unemployment is the unemployment rate factor. CPI is the consumer price index, which is a proxy for inflation factor. All four macroeconomics factors are based off OECD data for Euro area, and in surprise ratio (actual minus expected figure). Adj-R^2 represents the adjusted R-squared value for goodness of fit of the model to the data.

[8] Dow Jones Industrial Average

[9] S&P Global Flash Composite Purchasing Managers' Index

The macroeconomic factor of BOP, CPI and Short-Term Interest, does not seem to have any significance in the estimate. On the other hand, the unemployment factor does seem to explain for itself as an influence on the financial market. A 1% change in the unemployment factor surprise ratio leads to approximately 59% change in the returns of the portfolios. However, the model matches poorly to the data as the adjusted R^2 is between 0.06 to 0.07. To improve the goodness of fit, both the inflation and short-term interest rate factor were dropped. The alpha from panel A, for both portfolios, were like the previous model with 4 macroeconomic factors. It is positive and significant at 90% and 95% confidence for the dirty and clean portfolio, respectively. There is still no significance in the BOP factor while the results of the unemployment factor were like the coefficient estimates of panel B.

The adjusted R^2 of both panels A and B remains very low. Hence, both models did not help determine if a clean portfolio was more resilient than a dirty portfolio. The only relevant interpretation comes from just the summary statistics' quarterly returns. A clean portfolio has a higher mean return and lower volatility than a dirty portfolio.

7. Discussion

Finally, the four-part analysis results from chapter 6 will be pieced together for discussion here.

In the first hypothesis, it was hypothesised that there was no statistical significance in performance difference between clean and dirty portfolios. Conferring to tables 7 and 10 alphas, we can reject hypothesis 1. From table 7 CAPM model, the clean portfolio not only has significance in its differences in returns against the dirty portfolio, but its outperformance led to a higher and statistically significant alpha than the dirty portfolio. Moreover, a clean portfolio has statistical significance in its outperformance when regressed using the 3-factor model. That outperformance is slightly higher than the dirty portfolio's alpha as well. Moving onto table 10, we can also see that both the 3-factor and 4-factor models illustrate significant alpha for a clean portfolio, albeit there was no significance in the difference regressand. Lastly, the mean returns from tables 6 and 9 also support the clean portfolio's higher average returns relative to the dirty portfolio. Therefore, the initial expectation of the outcome of the first hypothesis turns out to be true. The results show statistical significance in performance difference between the clean and dirty portfolios and that the clean portfolio outperforms the dirty portfolio. The findings are aligned with Kempf and Osthoff (2007), where authors concluded that a portfolio which buys high ESG score firms and sells low ESG score firms led to 8.7% alpha per annum. Statman and Glushkov (2009) also shared similar results with their conclusion that a portfolio with stock tilts toward firms with high social responsibility while avoiding firms with low social responsibility can gain abnormal returns.

However, the results above ran contrary to the findings of Bauer et al. (2005), Cortez et al. (2012), and Halbritter and Dorleitner (2015). In these papers, the authors found no statistical significance in the performance difference between a sustainable portfolio to a conventional portfolio. Looking through the lens of SRI funds, although investments into (a portfolio of) SRI funds do have a significance in performance difference, several papers found that they underperformed instead when compared to conventional funds and benchmark markets (Girard et al. r, 2007; Renneboog et al., 2008; Lee et al., 2010; Kanuri, 2020; and Döttling & Kim 2022).

The second hypothesis concerning a clean portfolio having higher performance measures than a dirty portfolio is not rejected. This was proven by table 8, where the use of conventional and alternative risk components was taken. The results of table 8 are in line with the hypothesis. Based on the full volatility component, systematic risk component, LPM-based risk measure and drawdown-based risk measure, the clean portfolio performed better than the dirty portfolio 75% of the time. By better, it means the clean portfolio's performance measures are higher than those of the dirty portfolio. The findings from the performance measures were similar to Sabbhagi (2011), where green ETFs were reported to have positive returns and outperformed the S&P 500 index, which a conventional ETF can track. The author, however, does find that these green ETFs had higher volatility and underperformed the S&P 500 during the global financial crisis. When looking at sector-specific, Miralles-Quirós and

Miralles-Quirós (2019) examined the performance difference between portfolio ETFs of renewable energy and conventional energy, and their results were aligned with this paper's findings. The authors concluded that the portfolio of renewable energy ETFs significantly outperformed the portfolio of conventional energy ETFs returns. Lastly, Lagewaard (2020) examined the performance difference between socially responsible and conventional funds. On the former, the author concluded that socially responsible funds have a higher Jensen's alpha, similar to the findings of this paper's performance measures analysis.

Hence, when looking at the main research question, a clean portfolio does perform comparatively better than a dirty portfolio in terms of the outperformance and risk-adjusted returns of the performance measures. Its better performance is not only over the long run (as seen in part 1), but also it is less affected by COVID-19 (as seen in part 3). Comparing the results to other research, this finding is in line with Nofsinger and Varma (2014), which concluded that socially responsible investment (SRI) funds outperformed conventional funds during the dot-com and global financial crisis. Pastor and Vorsatz (2020) also support this finding with their research into mutual funds' performances during COVID-19. Four-fifth of the mutual funds showed underperformance during COVID-19, but those with better Morningstar sustainability scores had better performances during COVID-19.

Furthermore, the alphas of the 5-factor and 6-factor models showed no significance in the underperformance of a clean portfolio during a crisis. The findings are in line with Pavlova and de Boyrie (2021), where ESG ETFs performed as well as the US market during the COVID-19 crisis. Moreover, the results of studies into equities with higher ESG scores support the findings above and the summary statistics. Stocks with better ESG ratings earned relatively higher returns and lower volatilities (Ding et al., 2020; Albuquerque et al., 2020). Broadstock et al. (2020) have similar findings in the Chinese market. The high-ESG portfolio outperformed the low-ESG portfolio, and the ESG performances helped reduce portfolio risk during the crisis of COVID-19.

Contrary to the previous results, Kanuri (2020) examined ESG ETFs and concluded that they did not outperform during the global financial crisis. Furthermore, Folger-Laronde et al. (2020) studied ESG ETFs performances in Turkey. The ETFs with higher sustainability ratings were not shielded from severe losses during COVID-19. In addition, this paper's findings are at odds with Döttling and Kim (2022) research into mutual funds' flow during COVID-19. They concluded that SRI funds performed worst off, as evidenced by their sharper decline in fund flows.

What about how resilient both portfolios are regarding their factor loading sensitivities? From the third hypothesis, it was expected to be rejected, and the outcome was expected to have significance in terms of factor loading sensitivities. Based on the results of the coefficient estimates, the hypothesis is rejected, and the expectation turns out to be true. When looking at the dummy variable, the clean portfolio is more resilient than the dirty portfolio as it is less negative in its estimates—as seen from

CAPM, 3-factor and 4-factor models. i.e., a clean portfolio was less impacted negatively compared to having a dirty portfolio.

When it comes to the market factor, the results were inconclusive. On the one hand, the Fama-French 5-factor model and the 6-model showed the clean portfolio is significant and more sensitive to market movements—such as deterioration from COVID-19—relative to the dirty portfolio. On the other hand, the market factor estimates from CAPM, 3-factor and 4-factor models for clean portfolios, are now significant and less sensitive to dirty portfolios. The same can be said for the size factor, whereby in the 5-factor and 6-factor models, clean portfolios have significant and lower sensitivity to size effect relative to dirty portfolios. Regarding the 3-factor and 4-factor models, the clean portfolio's size estimate is now significant but has higher sensitivity than the dirty portfolio. The 3-factor and 4-factor model estimates align with the findings of Cortez et al. (2012) on SRI funds of Europe and the US, sharing small-cap bias. Furthermore, Bauer et al. (2005) concluded that in mutual funds, the small-cap bias is more substantial for European than US funds.

In a dated paper, Luther et al. (1992) shared that UK companies with higher sustainability ratings—ethical unit trusts—tend to be growth companies. Hence, these growth companies are more likely to be constituents in articles 8 and 9 ETFs than in article 6 ETFs. The value factor estimates for the differences from the 3-factor model in table 7 and Table 10 support this finding. The 5-factor model's value factor estimates for the differences also concur with the view. In these difference estimates, they were negative and significant at 90 to 95% confidence. i.e., value stocks' outperformance is inverse to the difference in returns for a clean portfolio over a dirty portfolio.

Besides momentum factor estimates from the 6-factor model, clean portfolio factor estimates for value, profitability, and investment policy factor have no significance. Therefore, only one interpretation can be derived from these factor estimates. Noting from those factor estimates with significance, they are between -1 to 1. One can infer clean portfolio is indeed resilient against those factor exposures. i.e., when the size or market factor moves by 1, even during COVID-19 periods, the clean portfolio returns move less than 1 unit, ceteris paribus. As to whether the clean portfolio is more resilient than the dirty portfolio for those factor exposures, valid interpretations cannot be made due to conflicting estimates.

Finally, in the fourth hypothesis, using macroeconomic factor analysis, it was hypothesised that there is no statistical significance for the estimates. This fourth hypothesis can be rejected, and the expectation of the outcome is in line with the hypothesis. The macroeconomic model showed one of the factors estimated to have significance—the unemployment rate. The clean portfolio and dirty portfolio returns are susceptible to unemployment rate surprises. Throughout the observed quarterly periods, all four macroeconomic indicators had a positive surprise ratio. i.e., the actual figures turn out to be higher than expected, leading to a positive surprise. What this meant for the unemployment rate, with its average surprise ratio of 0.3%, is that when the unemployment rate surprise factor moves by

1%, the clean portfolio returns were positively influenced by 49 to 59% on its returns. The significance of the unemployment rate estimates is similar to the findings of Boyd et al. (2015). In that paper, the author found significance in the relationship between the unemployment rate and the stock returns, but it depends on the state of the economy. During a contractionary economy, the relationship between unemployment and stock returns is inverse when the news announces a rising unemployment rate.

While the estimates for CPI showed no significance in relationship with the clean and dirty ETF portfolios returns, several papers found results contrary to this paper's findings. Flannery and Protopapadakis (2002) conclude that CPI correlates significantly with stock returns. Specifically, CPI was found to have a significant and inverse relationship with asset classes (Kolluri & Wahab, 2008). Moving on to the lack of significance between short-term interest rates and the portfolio returns, the results are similar to Çiftçi (2014). In that paper, the author found no significance in the influence of interest rates on stock returns.

To end, the macroeconomic factor model showed that a clean portfolio also outperforms a dirty portfolio in the long run with higher alpha. The period studied—45 quarters—is long-term and goes in line with the findings of multifactor models from parts 1 and 3 analysis. This shows that the clean portfolio of ETFs outperforms the dirty portfolio of ETFs from the short-term to the long-term investment horizon..

8. Conclusion

Considering the pressing need to transition from conventional investment strategies to one that is sustainable, the trend of sustainable investing is growing exponentially. The master's thesis aims to examine whether responsible investing comes at a performance cost and if responsible investing is resilient during the COVID-19 crisis. While there is a wealth of research examining the performances during crisis of sustainable equities and mutual funds, few can be found to have investigated exchange-traded funds—especially rare on the COVID-19 crisis. In addition, with a myriad of ESG rating agencies, asset classes have several non-cohesive ESG scores. The advent of SFDR helps to harmonise with one technical standard. As such, using the SFDR standard, the four methods were applied to ETF portfolio throughout the paper to shed new light on the main research question.

The clean ETF portfolio performed better than the European market based on the monthly returns regression analysis. Looking at the regression analysis of the COVID-19 event study, there are outperformances of the clean portfolio using the 3-factor and 4-factor models. Moreover, the performance measures further support the findings. For 75% of the time, the performance measures showed higher scores for a clean portfolio than a dirty portfolio. The macroeconomic factor analysis proves that a clean portfolio not only has better performances during COVID-19 and in the medium-term, but it also has outperformance over the long run of 45 quarters.

Regarding resilience, clean portfolio ETFs were less impacted during COVID-19, using the CAPM, 3-factor and 4-factor models. Moving on to the market factor, both ETF portfolios were less sensitive than the market factor using the 3-factor, 4-factor, 5-factor, and 6-factor models. This points to the benefits of diversification. However, comparing the estimates of clean and dirty portfolios led to inconclusive answers. In size factor, the factor estimates gave inconclusive answers as well.

On the one hand, 3-factor and 4-factor model shows more exposure of a clean portfolio to size factor tilts. On the other hand, 5-factor and 6-factor show the opposite. Meaningful interpretation of these factor loadings can still be drawn from the difference as a regressand. Both size and value factors were negative and significant in the coefficient estimate. The results show that a clean portfolio is more exposed to small-cap and growth stocks. There was no significance in the two quality factors of Fama-French and Carhart's momentum factor.

Coming back to the **main research question** of this paper:

What is the performance difference between a clean and a dirty portfolio, and how does the resilience differ during COVID-19 exogenous shock?

The findings of this master's thesis show that, yes, there is a performance difference between clean and dirty portfolios. Yes, the clean portfolio outperformed the dirty portfolio and was more resilient during COVID-19 by being less negatively impacted.

8.1 Limitations of Research

Identifying the research limitations is essential to understand to what extent findings represent the population of ETFs in Europe, rather than being restricted to the sample under examination. Four main limitations can be highlighted in this study.

Firstly, the sample size for sustainable ETFs limited. Article 8 ETFs for Europe, denominated in Euro, had the following sample size: 25 for daily returns, 17 for monthly returns and 10 for quarterly returns. Article 9 ETFs also had sample sizes between 2 to 4, depending on the return frequency. Both categories had to be combined to form a clean portfolio. The low number of sustainable ETFs has to do with the lack of sustainable ETF offerings and the lack of data availability. On the former, sustainable ETFs were only in vogue in recent years. Notice how there were more ETFs in the sample size for daily returns, which examined the period between 5 June 2019 to 5 November 2020. As seen in table 1, as the period under study stretches back to 2010 (quarterly returns), the number of sustainable ETFs dropped dramatically. This is due to the lack of indices in the past that catered to sustainable investment strategies.

As a consequence, asset managers cannot offer sustainable ETFs even if they want to do due to the absence of ESG indices and sustainable constituents to make them. Lack of data availability was also an issue during the data inspection. Several ETFs, both sustainable and non-sustainable, had missing returns even though the ETFs were already in existence. This meant that those ETFs had to be removed from the sample size as well.

Secondly, survivorship bias is also present. When looking at exchange-traded funds, any changes in the methodology of inclusion into a sustainable index can cause constituents to be dropped from the index. Regardless of the (non-)sustainability aspect, constituents, in general, can also be dropped due to poor performance or firm closure. Survivorship bias can cause overestimation in the performance of ETFs and result in inaccurate findings. The manual search for ETFs with articles 6, 8 and 9 classifications also made the list of ETFs non-exhaustive. Financial databases, such as Bloomberg and Morningstar, provided sustainability screening criteria based on proprietary ESG scores. Consequentially, sustainable ETFs with articles 8 or 9 could have been missed out despite the best attempt to search all major asset managers' fund prospectus.

Thirdly, the time-series dataset tends to have a serial correlation. Post estimation statistics from table 2 show the presence of heteroskedasticity and autocorrelation. Ignoring both leads to coefficient estimates, using OLS, to still be unbiased but inefficient. The standard error estimates could be wrong, and any inferences made can be misleading. For the presence of heteroskedasticity, log transformation of the factor loadings was not feasible as the logarithms of a variable can only occur for positive variables. i.e., non-zero and not negative values.

Moving on to the presence of autocorrelation, one approach to dealing with this violation is to move the model specification from static to dynamic using lagged variables. However, including lagged variables in the regressors or regressand brings additional problems. Namely, lagged regressand will violate the assumption of regressors being non-stochastic. This is because, by definition, the regressor is partly explained by the error term. Hence, the lagged value cannot be non-stochastic (Brooks, 2019). Adding lags regressor may solve serial correlation in residual, but it will come at the expense of creating an interpretational one. Brooks (2019) concluded that a large number of lags makes it challenging to interpret and test the hypothesis, which required the use of regression analysis in the beginning.

Hence, the Newey-West estimator was used in the time-series regression analysis to deal with heteroskedasticity and autocorrelation. No doubt, the use of it will also cause the standard errors to be more significant than they should have been with OLS standard errors. The regression analysis will be made more conservative.

Lastly, the macroeconomic factor model in part 4 was used as it was intuitive to interpret and link it with macroeconomic theories. However, it proved useless as the model did not fit the data well. The low adjusted R^2 meant that the model had low explanatory power. This was because the factor surprise ratios were used instead of the macroeconomic data. Connor (1995) concluded that the low explanatory power was due to the need to identify and quantify all pervasive shocks that impacted the underlying stock returns in the indices that the ETFs track. In this case, the four macroeconomic factors do not capture all of the impacts, and more factor surprises were needed.

8.2 Recommendations for Further Research

First, the sample size of ETFs, especially articles 8 and 9, can be increased by including non-Euro-denominated ETFs based on Europe—such as Pound Sterling or Swiss Franc ETFs. Next, readers of this thesis with econometric knowledge can use panel data regression instead of time-series regression. When examining performance differences, pooled OLS or fixed effects models can be used. The model treats the dataset as cross-sectional data for the former and ignores the time dimensional aspect. In the latter, the model takes it one step further by accounting for differences between the different currency denominations. For instance, the regressors will have a fixed effect on Pound Sterling and Swiss Franc.

Second, when it comes to the presence of autocorrelation and heteroskedasticity, future research can use the generalised least square (GLS) model instead of the OLS model. The Cochrane-Orcutt procedure can be conducted to treat for autocorrelation in the residuals before running GLS regression (Cochrane & Orcutt, 1949).

Third, in replacement of the macroeconomic factor model, Connor (1995) concluded that the fundamental or statistical factor model offers higher explanatory power. The former will be recommended as it marginally outperforms the statistical factor model (Connor, 1995). The fundamental factor model can be conducted by examining the underlying constituents of ETFs for their fundamentals, like dividend yields. The multifactor model from Fama-French, such as the value factor, is one good example of a fundamental factor model using book-to-market ratio.

8.3 Concluding Remarks

To conclude this thesis, the year 2022 has been marked with upheavals, from an exceptionally high inflationary environment to the energy crisis worldwide. This presents another opportunity to conduct studies into performance resilience for responsible investment. For instance, will a clean portfolio of sustainable ETFs serve as a haven against interest rate risk? Central banks globally have been aggressive with interest rate hikes to clamp down inflation. Today, 27 July 2022, Jerome Powell from the Fed announced another round of 75bps rate hikes, making recession this year ever more likely. Consequentially, further studies such as this master's thesis are made ever more pressing and indispensable. Thank you.

9. References

Adler, T., & Kritzman, M. (2008). The cost of socially responsible investing. *The Journal of Portfolio Management*, *35*(1), 52–56. https://doi.org/10.3905/jpm.2008.35.1.52

Albuquerque, R., Koskinen, Y., Yang, S., & Zhang, C. (2020). Resiliency of environmental and social stocks: An analysis of the exogenous COVID-19 market crash. *The Review of Corporate Finance Studies*, 9(3), 593–621. https://doi.org/10.1093/rcfs/cfaa011

Allen, M. P. (1997). The Problem of Multicollinearity. *Understanding Regression Analysis* (pp. 176–180). Boston, MA: Springer. https://doi.org/10.1007/978-0-585-25657-3_37

Auer, B. R. and Schuhmacher, F. (2016). Do socially (ir)responsible investments pay? new evidence from international ESG data. *The Quarterly Review of Economics and Finance*, 59:51–62

Bauer, R., Koedijk, K., & Otten, R. (2005). International evidence on Ethical Mutual Fund Performance and Investment Style. *Journal of Banking & Finance*, 29(7), 1751–1767. https://doi.org/10.1016/j.jbankfin.2004.06.035

Bauer, R., Ruof, T., & Smeets, P. (2021). Get real! individuals prefer more sustainable investments. *The Review of Financial Studies*, *34*(8), 3976–4043. https://doi.org/10.1093/rfs/hhab037

Ben-David, I., Li, J., Rossi, A., & Song, Y. (2021). What do mutual fund investors really care about? *The Review of Financial Studies*, *35*(4), 1723–1774. https://doi.org/10.1093/rfs/hhab081

Berk, J., & van Binsbergen, J. (2021). The Impact of Impact Investing. *Stanford University Graduate School of Business Research Paper*. http://dx.doi.org/10.2139/ssrn.3909166

Blitz, D., Hanauer, M. X., Vidojevic, M., & van Vliet, P. (2016). Five concerns with the five-factor model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2862317

Boyd, J. H., Hu, J., & Jagannathan, R. (2005). The stock market's reaction to unemployment news: Why bad news is usually good for stocks. *The Journal of Finance*, *60*(2), 649–672. https://doi.org/10.1111/j.1540-6261.2005.00742.x

Brammer, S. J., Brooks, C., & Pavelin, S. (2005). Corporate Social Performance and Stock Returns: UK evidence from Disaggregate Measures. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.739587

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models*. *Australian Economic Papers*, *17*(31), 334–355. https://doi.org/10.1111/j.1467-8454.1978.tb00635.x

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287. https://doi.org/10.2307/1911963

Broadstock, D., Chan, K., Cheng, L. T., & Wang, X. W. (2020). The role of ESG performance during times of financial crisis: Evidence from covid-19 in China. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3627439

Brooks, C. (2019). *Introductory Econometrics for Finance* (4th ed.). Cambridge: Cambridge University Press. doi:10.1017/9781108524872

Burke, G., 1994. A sharper Sharpe ratio. *Futures,* 23 (3), 56.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance, 52* (1), 57-82.

Çiftçi*,* S. (2014) *The influence of macroeconomic variables on stock performance* [Master's thesis, University Twente]. University of Twente Thesis Repository. https://essay.utwente.nl/66492/1/S.Ciftci_MA_Business%20Administration.pdf

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto- correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32. https://doi.org/10.2307/2280349

Connor, G. (1995). The Three Types of Factor Models: A Comparison of Their Explanatory Power. *Financial Analysts Journal*, *51*(3), 42–46. http://www.jstor.org/stable/4479845

Cortez, M. C., Silva, F., & Areal, N. (2012). Socially responsible investing in the global market: Socially responsible investing in the global market: The performance of US and European funds. *International Journal of Finance & Economics*, 17(3), 254–271. https://doi.org/10.1002/ijfe.454

Del Guercio, D., & Tkac, P. A. (2008). Star Power: The effect of Morningstar ratings on mutual fund flows. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.286157

Ding, W., Levine, R., Lin, C., & Xie, W. (2020). Corporate immunity to the COVID-19 pandemic. *NBER*. https://doi.org/10.3386/w27055

Döttling, R., & Kim, S. (2020). Sustainability preferences under stress: Evidence from mutual fund flows during COVID-19. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3656756

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*(3/4), 409. https://doi.org/10.2307/2332391

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression. II. *Biometrika*, *38*(1/2), 159. https://doi.org/10.2307/2332325

Dybvig, P. H., & Ross, S. A. (1985). Differential information and performance measurement using a security market line. *The Journal of finance*, *40*(2), 383-399.

Eling, M., & Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking & Finance*, 31(9), 2632–2647. https://doi.org/10.1016/j.jbankfin.2006.09.015

Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal Of Financial Economics*, *33*(1), 3-56. doi: 10.1016/0304-405x(93)90023-5

Fama, Eugene F. & French, Kenneth R., 2015. "A five-factor asset pricing model," *Journal of Financial Economics*, Elsevier, vol. 116(1), pages 1-22.

Fama, Eugene, F., and Kenneth R. French. 2004. "The Capital Asset Pricing Model: Theory and Evidence." *Journal of Economic Perspectives*, 18 (3): 25-46. https://doi.org/10.1257/0895330042162430

Ferrell, A., Liang, H., & Renneboog, L. (2014). Socially responsible firms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2473502

Flannery, M. J., & Protopapadakis, A. A. (2002). Macroeconomic factors do influence aggregate stock returns. *Review of Financial Studies*, *15*(3), 751–782. https://doi.org/10.1093/rfs/15.3.751

Folger-Laronde, Z., Pashang, S., Feor, L., & ElAlfy, A. (2020). ESG ratings and financial performance of exchange-traded funds during the COVID-19 pandemic. *Journal of Sustainable Finance & Investment*, *12*(2), 490–496. https://doi.org/10.1080/20430795.2020.1782814

Gantchev, N., Giannetti, M., & Li, Q. (2020). Sustainability or performance? ratings and fund managers' incentives. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3731006

Girard, E. C., Rahman, H., & Stone, B. A. (2007). Socially responsible investments. *The Journal of Investing*, *16*(1), 96–110. https://doi.org/10.3905/joi.2007.681827

Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, *46*(6), 1293. https://doi.org/10.2307/1913829

Guo, D., & Zhou, P. (2021). Green bonds as hedging assets before and after COVID: A comparative study between the US and China. *Energy Economics*, 104, 105696. https://doi.org/10.1016/j.eneco.2021.105696

Halbritter, G. and Dorfleitner, G. (2015). The wages of social responsibility—where are they? a critical review of ESG investing. *Review of Financial Economics*, 26:25–35.

Hale, J. (2021). Sustainable equity funds outperform traditional peers in 2020. Morningstar Collection. Retrieved from https://www.morningstar.com/articles/1017056/sustainable-equity-funds-outperform-traditional-peers-in-2020

Henisz, W. J., Dorobantu, S., & Nartey, L. J. (2013). Spinning gold: The financial returns to stakeholder engagement. Strategic Management Journal, 35(12), 1727–1748. https://doi.org/10.1002/smj.2180

Hillman, A. J., & Keim, G. D. (2001). Shareholder value, stakeholder management, and social issues: What's The bottom line? *Strategic Management Journal*, *22*(2), 125–139. https://doi.org/10.1002/1097-0266(200101)22:2<125::aid-smj150>3.0.co;2-h

Hong, H., and Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics*, 93(1), 15-36.

Horstmeyer, D., Liu, Y., & Wilkins, A. (2022). Fama and French: The Five-Factor Model Revisited. *CFA Institute*. Retrieved July 21, 2022, from https://blogs.cfainstitute.org/investor/2022/01/10/fama-and-french-the-five-factor-model-revisited/

Jensen, M., 1968. The performance of mutual funds in the period 1945–1968. *Journal of Finance,* 23 (2), 389–416.

Kanuri, S. (2020). Risk and return characteristics of environmental, social, and governance (ESG) equity etfs. *The Journal of Index Investing*, *11*(2), 66–75. https://doi.org/10.3905/jii.2020.1.092

Kempf, A. and Osthoff, P. (2007). The effect of socially responsible investing on portfolio performance. *European Financial Management*, 13(5):908–922.

Kestner, L.N., 1996. Getting a handle on true performance. *Future,s* 25 (1), 44–46.

Kolluri, B., & Wahab, M. (2008). Stock returns and expected inflation: Evidence from an asymmetric test specification. *Review of Quantitative Finance and Accounting*, *30*(4), 371–395. https://doi.org/10.1007/s11156-007-0060-9

Lagewaard, J. (2020) *Does it pay to be good? The effect of socially responsible investing on the performance of mutual funds* [Bachelor's thesis, Erasmus University Rotterdam]. Erasmus University Thesis Repository. https://thesis.eur.nl/pub/53403/Thesis-Bachelor-IBEB.pdf

Lee, D. D., Humphrey, J. E., Benson, K. L., & Ahn, J. Y. (2010). Socially responsible investment fund performance: the impact of screening intensity. *Accounting & Finance*, 50(2), 351-370.

Lins, K. V., Servaes, H., & Tamayo, A. (2017). Social Capital, Trust, and firm performance: The value of corporate social responsibility during the financial crisis. *The Journal of Finance*, *72*(4), 1785–1824. https://doi.org/10.1111/jofi.12505

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, 47(1), 13–37. https://doi.org/10.2307/1924119

Luther, R. G., Matatko, J., & Corner, D. C. (1992). The investment performance of UK "ethical" unit trusts. *Accounting, Auditing & Accountability Journal*, *5*(4). https://doi.org/10.1108/09513579210019521

Markowitz, H. (1959). Portfolio selection: efficient diversification of investments (Ser. Cowles foundation for research in economics. monograph, 16). *Yale University Press*.

Miralles-Quirós, J. L., & Miralles-Quirós,M.M.(2019).Are alternative energies a real alternative for investors? *Energy Economics*, 78, 535–545.

Modigliani, F., Modigliani, L., 1997. Risk-adjusted performance – how to measure it and why. *Journal of Portfolio Management,* 23 (2), 45–54.

Morgan Stanley Institute for Sustainable Investing. (2019). *Sustainable Reality. Analyzing Risk and Returns of Sustainable Funds.* Retrieved from https://www.morganstanley.com/content/dam/msdotcom/ideas/sustainable-investing-offers-financial-performance-lowered risk/Sustainable_Reality_Analyzing_Risk_and_Returns_of_Sustainable_Funds.pdf

Nofsinger, J., & Varma, A. (2014). Socially responsible funds and market crises. *Journal of Banking & Finance*, 48, 180–193. https://doi.org/10.1016/j.jbankfin.2013.12.016

Pastor, L., & Vorsatz, M. B. (2020). Mutual Fund performance and flows during the COVID-19 crisis. *NBER.* https://doi.org/10.3386/w27551

Pavlova, I., & de Boyrie, M. E. (2022). ESG ETFs and the COVID-19 stock market crash of 2020: Did clean funds fare better? *Finance Research Letters*, *44*, 102051. https://doi.org/10.1016/j.frl.2021.102051

Plackett, R. L. (1949). A Historical Note on the Method of Least Squares. Biometrika, 36(3/4), 458–460. https://doi.org/10.2307/2332682

Pokorna, M. (2017) *Socially Responsible Investing and Portfolio Performance* [Master's thesis, Erasmus University Rotterdam]. Erasmus University Thesis Repository. https://thesis.eur.nl/pub/41396/Pokorna-M.-449106-.pdf

Renneboog, L., ter Horst, J. R., & Zhang, C. (2008). The price of ethics and stakeholder governance: The performance of Socially Responsible Mutual Funds. *SSRN Electronic Journal*. https://doi.org/10.1016/j.jcorpfin.2008.03.009

Roll, R. (1978). Ambiguity when performance is measured by the securities market line. *The Journal of Finance*, 33(4), 1051-1069.

Sabbaghi, O. (2011). Do Green Exchange-Traded Funds Outperform the S&P500. *Journal of Accounting and Finance*, 11(1), 50–59.

Schnietz, K. E., & Epstein, M. J. (2005). Exploring the financial value of a reputation for corporate social responsibility during a crisis. *Corporate Reputation Review*, *7*(4), 327–345. https://doi.org/10.1057/palgrave.crr.1540230

Servaes, H., & Tamayo, A. (2013). The impact of corporate social responsibility on firm value: The role of Customer Awareness. *Management Science*, *59*(5), 1045–1061. https://doi.org/10.1287/mnsc.1120.1630

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The *Journal of Finance*, 19(3), 425. https://doi.org/10.2307/2977928

Sharpe, W.F. (1966) Mutual fund performance. *The Journal of Business*, 39, 119-138. https://doi.org/10.1086/294846

Sortino, F.A., van der Meer, R., 1991. Downside risk. *Journal of Portfolio Management,* 17 (Spring), 27–31.

Sortino, F.A., van der Meer, R., Plantinga, A., 1999. The Dutch triangle. *Journal of Portfolio Management,* 26 (Fall), 50–58.

Statman, M. and Glushkov, D. (2009). The wages of social responsibility. *Financial Analysts Journal*, 65(4):33–46.

Tangjitprom, N. (2012). The review of macroeconomic factors and stock returns. *International Business Research*, *5*(8). https://doi.org/10.5539/ibr.v5n8p107

Treynor, J. L. (1961). Market value, time, and risk. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2600356

Treynor, J.L., 1965. How to rate management of investment funds. *Harvard Business Review,* 43 (1), 63–75.

US SIF Foundation. (2020). *Report on US sustainable and impact investing trends*. Retrieved May 4, 2022, from https://www.ussif.org/files/US%20SIF%20Trends%20Report%202020%20Executive%20Summary.pdf

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817. https://doi.org/10.2307/1912934

Young, T.W., 1991. Calmar ratio: A smoother tool. *Futures,* 20 (1), 40.

10.Appendices

Appendix A. Supportive literature of sustainability performance difference and resilience

Table 13. Overview of literature supportive of responsible investing

Year	Paper	Research Question	Data Sample	Data Period	Main Finding
2001	Hillman & Keim	Does a relationship exist between shareholder value, stakeholder management and social issues.	308 firms	1994 to 1996	Concerning ESG factors, stakeholder management (G factor) increases shareholder value and firm's competitive advantage. Using firm resources to tackle social issues (S factor) that do not concern shareholders does not lead to shareholder value creation.
2005	Schnietz & Epstein	Does firm reputation for CSR translate into financial value during crisis.	416 firms from Fortune 500	26 Nov 1999 to 29 Nov 1999	Reputation for CSR protected firms from stock declines associated with 1999 Seattle World Trade Organisation crisis, even after controlling for possible trade and industrial effects.
2007	Kempf & Osthoff	Does incorporating SRI screens into investment process increase portfolio performances.	All stocks of S&P 500 and DS 400	1992 to 2004	Based on KLD SRI ratings, buying stocks with high SRI score and selling stocks with low SRI score leads to 8.7% abnormal returns per annum. Highest returns are achieved when several screens are combined at the same time, to buy stocks with extreme SRI ratings. These returns remain significant after accounting for transaction costs.
2009	Statman & Glushkov	What is the impact on performance for tilt towards stocks with high social responsibility and shunning stocks with low social responsibility.	2955 companies from KLD database	1992 to 2007	For socially responsible portfolios, the advantage from the tilt toward stocks with high social responsibility (higher abnormal returns against conventional peers) is offset by the disadvantage from excluding stocks of low social responsibility. Hence, the net effect is zero. Paper concludes that SR investors can gain abnormal returns by continuing with stock tilts towards high SR score but refrain from excluding stocks with low SR score in portfolio.
2013	Henisz et al.	What is the relationship between stakeholder engagement and financial returns.	19 mining firms traded on Toronto stock exchange, Canada	1993 to 2008	In relation to the governance and social factor, increasing stakeholder engagement by firms do enhances financial valuation of firms. These 19 firms own 26 gold mines in 20 countries.
2013	Servaes & Tamayo	Can CSR activities from firm create shareholder value.	2000 firms from KLD Stats	1991 to 2005	CSR activities from firms does create value for shareholders but only under certain conditions. Conditions are that the firm has high public awareness proxied by advertising intensity, and consistency between firm's CSR efforts and overall reputation.
2014	Ferrell et al.	Does CSR have a relationship with agency problem. And does it lead to shareholder value destruction.	4700 large firms from 60 countries	2002 to 2013	Companies with higher CSR ratings have fewer agency problems, such as lower level of free cashflows, higher dividend payout, and higher leverage ratio. Relating to this paper, the relevant findings conclude CSR ratings are associated with maximising shareholder value.

2014	Nofsinger & Varma	Examined whether socially responsible (SR) mutual funds outperformed conventional peers during market crises. In addition, whether SR funds offer downside protection for investors during market crises.	240 US domestic equity mutual funds	2000 to 2022	SR funds outperformed (by 1.61-1.70% annualised) their matched conventional peers during market crises, providing downside protection for investors. However, SR funds comes at a cost for underperformance during non-crisis periods compared to conventional peers. The latter outperformed by annualised 0.67-0.95%.
2017	Lins et al.	Does firm with high social capital (CSR) have higher stock returns than low CSR firms during global financial crisis.	1673 nonfinancial firms	CSR data of firms taken at Dec 2006; stock returns of firm from Aug 2008 to Mar 2009	Firm with high social capital had 4-7% higher stock returns than firms with low social capital during the 2008-2009 financial crisis. Comparatively, high CSR firms exhibited higher profitability and growth rate. This paper evidenced how raising the level of trust with stakeholders, creating higher social capital, pays off in times of negative shocks to market.
2020	Albuquerque et al.	Are there any resilience in environmental (E) and social (S) stocks during COVID-19 market crash.	2123 US stocks	2017 to 2019	Stocks with high ES ratings have significantly higher returns, lower volatility, and higher profit margins during the crash.
2020	Broadstock et al.	Role of ESG performances during COVID-19 in China, and interpreting ESF performance as signal of future stock performances and/or risk mitigation.	300 Mainland China stocks from CSI300	22 Jan 2020 to 5 Feb 2020	Using constituents of China's CSI300, high ESG portfolios have tendency to outperform low ESG portfolio. ESG performance is positiviely associated with short-term cumulative returns of CSI300 stocks during COVID-19. ESG performance also helps to mitigate financial risk from financial crisis. ESG performances computed based on a variety of E, S and G indicators.
2020	Ding et al.	Examine the relationship between pre-2020 corporate characteristics and stock price reactions to the COVID-19 pandemic.	6000 firms across 56 countries	2018 to Q1 2020	The five corporate characteristics are financial conditions, international supply chain and customer locations, CSR intensity, corporate governance and owership structures. Relevant findings to this paper, firms with stronger financial conditions, higher CSR intensity or better governance structure prior to pandemic experienced superior stock price reaction to COVID-19 than others.
2020	Lagewaard	What are the effects of SRI on mutual funds' performance.	274 SR funds and 31633 conventional funds	2010 to 2019	Conventional funds have significance in lower excess return comparatively. Socially responsible funds have lower systematic risk relative to conventional funds.
2020	Pastor and Vorsatz	What is the effect of sustainability ratings on US active equity mutual funds performance during COVID-19 crisis	3626 US actively equity mutual funds	1 Jan 2017 to 30 April 2020	Most active funds underperformed passive benchmarks during COVID-19.US equity funds with better sustainability ratings and more star ratings from Morningstar performed better. These funds outperformed funds with lower sustainability and star ratings, against the FTSE,Russell benchmark. Investors view sustainability as necessity and not luxury during major crisis, hence they favour funds with high sustainability ratings when reallocating their capital.

2021	Guo & Zhou	Can green bonds serve as hedging asset during COVID-19	Barclays MSCI Green Bond Index (US); Essential Green Bond Index CNGB (China)	Aug 2014 to Aug 2021 (US); Jul 2014 to Aug 2021 (China)	Green bonds in US and China hedging effectiveness does decline during COVID-19, but the hedging effect is still positive. Green bonds hedging effects tends to be smaller for China than US pre-covid, but the gap has diminished since COVID-19. Thus, green bonds serve as a haven against financial shocks, not just COVID-19.
2021	Hale	How did sustainable equity funds fare against conventional peers in 2020.	26 ESG equity index funds	2020	25 out of 26 Sustainable index funds outperformed conventional peers in 2020. 11 out of 12 US large cap sustainable fund outperformed S&P 500. Outside of US, only 3 out of 11 index funds focusing on developed markets outperformed MSCI EAFE Index.
2021	Pavlova and de Boyrie	Examined if ESG ETFs fared better during COVID-19 market crash.	62 ESG ETFs	14 Nov 2019 to 29 May 2020.	Pre-COVID-19 crash, higher rated sustainable ETFs underperformed the market and their lower rated peers. During COVID-19 market crash, the higher rated ESG ETFs were not shielded from losses, but it did not perform worse than the market.

Appendix B. Opposing literature of sustainability performance difference and resilience

Table 14. Overview of literature opposing responsible investing

Year	Paper	Research Question	Data Sample	Data Period	Main Finding
1992	Luther et al.	Investment performance of UK ethical unit trusts	15 ethical unit trusts	Dec 1972 to Jun 1990	These ethical unit trusts have weak evidence of overperformance. Clear evidence of them skewed towards UK small cap than market. Also, tendency to invest in low dividend yield companies (growth stocks)
2005	Bauer et al.	Whether ethical mutual funds differ in terms of risk-adjusted return and investment style from matched sample of conventional funds.	103 German, UK and US ethical mutual funds	1990 to 2001	No significance in performance differences between ethical and conventional funds. Comparatively, ethical funds are less exposed to market return variability. UK, German ethical funds are heavily exposed to small caps, while US is exposed to large caps. Ethical funds in general exhibits growth stock tilts and less value tilts.
2006	Brammer et al.	Examine relationship between CSR factors and stock returns in UK. Investigation conducted at firm level and not fund level.	All constituents of FTSE All-Share Index (451 firms)	Variables based on 1 Jul 2002.	UK firms with higher social scores tend to have lower returns, vice versa.
2007	Girard et al.	Are SRI funds beneficial or do they come at a cost.	117 US SRI mutual funds	1984 to 2003	SRI fund managers show poor selectivity and market timing ability compared to Lippers active benchmark indices. SRI funds suffers from lack of diversification cost, older funds suffers more from this cost and even poor selectivity. More ethical screens contributes to these drawbacks.
2008	Adler & Kritzman	Does SRI screenings come at a cost.	Random sample of 500 SRI funds returns.	2007	SRI comes at a cost and it depends on manager skill and strictness of defining SRI. Cost is the highest for highly skilled managers who applied the stringent screening.
2008	Renneboog et al.	Does investors pay a price for SRI investments or do they benefit from superior returns.	440 SRI funds and 16036 non-SRI funds globally	1991 to 1995 (pre-bubble); 1996 to 1999 (internet bubble); 2000 to 2003 (post-bubble)	SRI funds in UK, UK, Europe and APAC underperformed domestic benchmark by -2.2% to -6.5%. Except for a limited number of countries like Japan, France and Sweden, SRI funds have no significant performance difference from conventional funds. Governance and social factor screns lead to lower returns.
2010	Lee et al.	Are there any benefits from investing in SRI funds, and does screening intensity impact the returns performance.	61 US Equity Funds	1989 to 2006	Non-financial screens comes at a cost to performance. The investment universe is constrained, diversification efficiencies decreased. Significance in reduction of 70bps in alpha per screen using Carhart model.
2012	Cortez et al.	Investigates the style and performance of US, European SRI funds	39 European SRI funds and 7 US SRI funds.	Aug 1996 to Aug 2008	Most European SRI funds have no significance in performance difference against conventional funds and socially responsible benchmarks. On the other hand, US and Austrian SRI funds comparatively show significance of underperformance.

2015	Halbritter & Dorleitner	Investigate the link between corporate social and financial performances of US firms, based on ESG.	4209 US companies (KLD); 1170 US firms (ASSET4); 1073 FIRMS (Bloomberg). 996 firms in common with all 3 databases.	2002 to 2011 (ASSET4); 2005 to 2011 (Bloomberg); 1990 to 2011 (KLD)	ESG portfolios have no significance in performance differences between companies with high ESG and those with low ESG. There is no significance in relationship between financial returns and ESG ratings based on 850 ESG indicators.
2016	Auer & Schuhmacher	Does SRI investments in APAC, US and Europe payoff.	632 APAC frms, 914 US firms, and 572 European firms.	Aug 2004 to Dec 2012	Active screening does not provide superior returns compared to passive investing. In US and APAC, SRI strategies do not underperform market. In Europe, SRI comes at a cost to performance.
2020	Folger-Laronde et al.	Are there differences and links between financial returns of ETFs and their sustainability performance during COVID-19 market downturn.	278 ETFs	11 Jan 2019 to 3 Mar 2020	This is an event study into COVID-19. The study found no link between sustainability and their financial performance. The sustainability also has no significance in resilience during market downturn. Higher sustainability does not safeguard against losses from severe market downturn.
2020	Kanuri	Investigate whether ESG ETF investing produce outperformance against equity market.	Undefined. Full paper not free to read.	Feb 2005 to Jul 2019	ESG ETF portfolio underperformed US (Russel 3000 ETF-IWC) and global (SPDR Global DOW ETF-DGT) equity markets.
2022	Döttling and Kim	Examined for the effect of COVID-19 on SRI mutual funds flow of retail and institutional funds	2720 retail funds, 2421 institutional funds	4 Jan 2020 to 25 Apr 2020	During COVID-19, funds with higher sustainability score had sharper decline in retail flows than institutional flows. SRI funds are exposed to higher sensitivities to retail flows during market shocks like COVID-19.