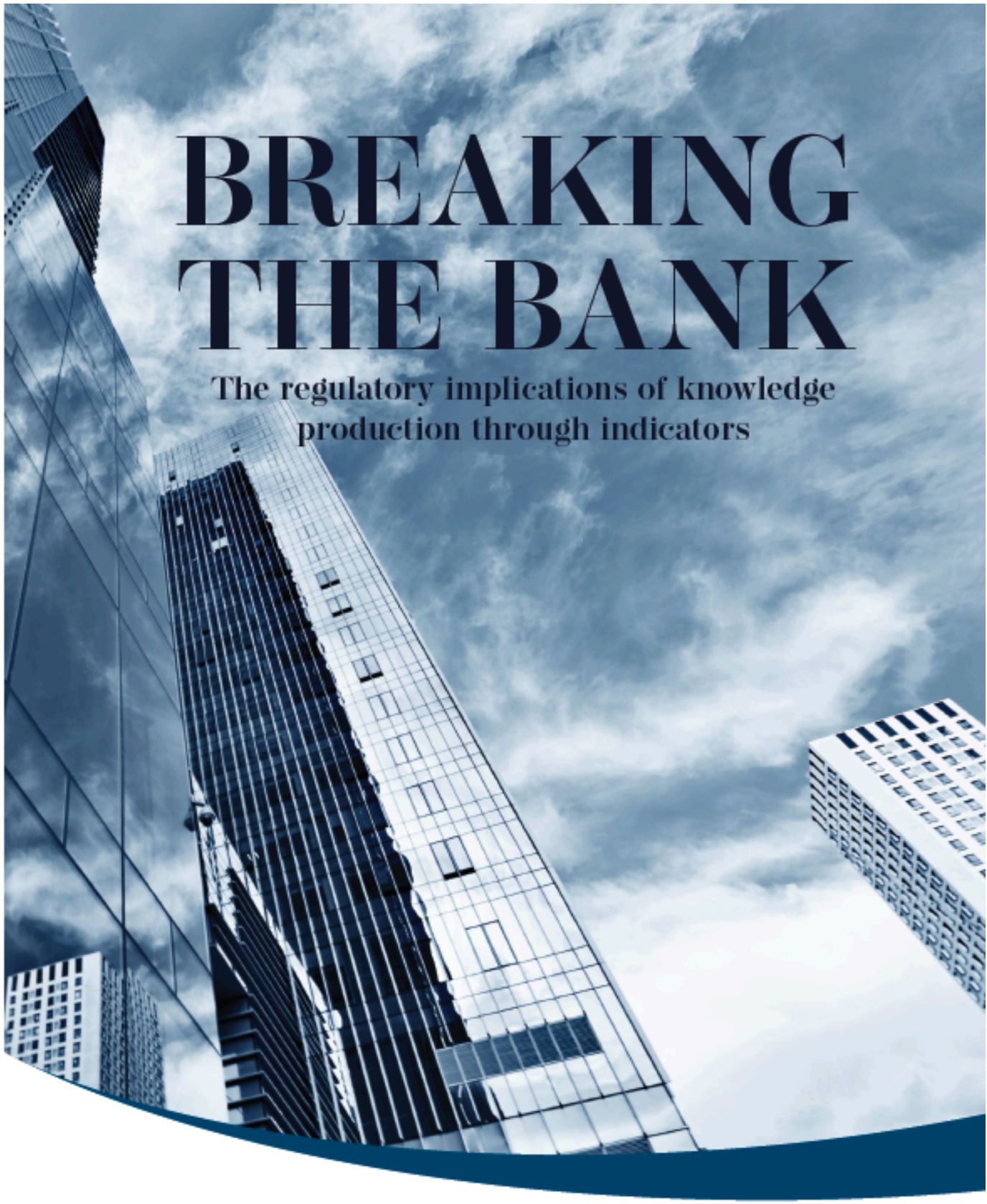

BREAKING THE BANK

The regulatory implications of knowledge production through indicators

Shirley Kempeneer

Universiteit Antwerpen

Faculty of Social Sciences

Political Science

Breaking the bank?

The regulatory implications of knowledge production through indicators

Proefschrift voorgelegd tot het behalen van de graad van doctor in de Sociale Wetenschappen: Politieke Wetenschappen aan de Universiteit Antwerpen

Te verdedigen door

Shirley KEMPENEER.

Promotor: Prof. dr. Wouter Van Dooren

Antwerpen 2019

Members of the Doctoral Committee:

Prof. Dr. Koen Verhoest

Prof. Dr. Gert Verschraegen

Members of the Doctoral Jury:

Prof. Dr. Veronica Junjan

Prof. Dr. Peter Lindner

This research was funded by the Research Foundation-Flanders (FWO) through a PhD fellowship grant (11X9818N).

Cover Picture credits by Hannah Declerck

TABLE OF CONTENTS

LIST OF TABLES, FIGURES, AND ABBREVIATIONS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANT	Actor-Network Theory
EBA	European Banking Authority
ECB	European Central Bank
JST	Joint Supervisory Team
LPF	Level Playing Field
NBB	National Bank of Belgium
NCA	National Competent Authority
NPM	New Public Management
PA	Public Administration
QA	Quality Assurance
SREP	Supervisory Review and Evaluation Process
ST	Stress Test
STS	Science and Technology Studies

PREFACE

Nothing we do, however virtuous, can be accomplished alone. I have been lucky enough to share this journey with brilliant and compassionate people, to whom I am forever grateful.

First of all, Wouter, I really enjoyed working and teaching with you. Thank you for having faith in me and the project. Your enthusiasm, advice, and support always kept me going. Koen, thanks for all your insightful comments over the years, and thanks for always creating a fun atmosphere in the research group. Gert, thank you for pushing me and always giving me new books and articles to read. Peter, thanks for having me in Frankfurt and giving valuable feedback on the project. Veronica, thanks for being in my jury, I admire you as an academic.

It has been said before, and it should be said again: The colleagues at the PW department are the best. Spending lonely days and nights pursuing a transcendent truth that only six or seven people will ever care about, is not always the most fun of jobs. But somehow, thanks to all of you, I look back on this experience with a smile on my face.

A couple of people deserve an extra word of thanks: First of all the AAP-gang, new and old, you guys are heroes for writing a PhD, and dealing with whiny and unappreciative students, I enjoyed being on your team; The VABAP-crew, and especially Jasmine, Dieter, and Sanne, thanks for taking my mind off my PhD; Patrick, thanks for trying to crash parties with me; Aydin, thanks for philosophising with me; Inger (De Wilde) thanks for all the advice; Inger (Baller) thanks for helping me with literally everything; Evelien, thanks for being the life of the party; Babette, thanks for always being there through the ups and downs; Brecht, thanks for being the best desk-buddy ever; Eva, thanks for being my interpretive soulmate.

Also outside of the Meerminne, there are a couple of people that kept me sane: The rest of the dinner crew, thanks for all the laughs; Julie, Charlotte, and Sarah, thanks for the much needed awesome and inappropriate breaks; Thy and Dana, thanks for interdisciplinary academic ranting over brunch; Monique, Melissa, and Nathalie, thanks for always being there with encouraging words and rosé; Jens, Stef, Wouter, and Max, thanks for singing my sorrows away with me; Eddie, thanks for the unconditional love; Kempeneer-crew, thanks for being such amazing role-models to your little sister, and thanks mom for always reminding me of what is really important in life.

Lauren, I would have never started this journey without you, and I could not have finished it either. From helping me write a research proposal in the middle of nowhere, to meticulously spell checking my final draft. You will always be my person.

Last but certainly not least, thanks are due to Frederik. You never cease to challenge me; academically and beyond. These chapters would not have been the same without you – and neither would I. I love you.

For Oma,

For everything.

INTRODUCTION

In 2008, Europe and the United States of America witnessed the worst economic downturn in the post-war era. Big ("too big to fail") banks were massively bailed out with public funds, as unemployment spiked and stock and home values plummeted (Geithner, 2015). In order to prevent such a crisis from happening again, governments in the United States and Europe introduced banking stress tests; an indicator to monitor the health of the banking system. In Europe, supervisory institutions[1] have conducted these stress tests on average every two years. Simply put, the stress test projects how much capital (measured as their risk-weighted capital ratio[2]) banks would have left after a three-year hypothetical crisis (adverse) scenario. An excerpt of the 2018 stress test results is presented in table 1 below (EBA, 2018).

TABLE 1: STRESS TEST RESULTS FOR INDIVIDUAL BANKS

Country	Bank	Adverse 2020
AT	Raiffeisen Bank AG	9.73%
AT	Erste Group Bank AG	8.56%
BE	KBC Group NV	13.60%
BE	Belfius Banque SA	13.21%
DE	Norddeutsche Landesbank	7.07%
DE	NRW.BANK	33.96%

[1] First the Committee of European Banking Supervision (CEBS) and after its abolishment in 2011 the European Banking Authority (EBA) in close cooperation with the European Central Bank (ECB).

[2] This is a banks' capital divided by its risk weighted assets. The idea of adding risk weights is that banks are required to hold more capital against risky assets. This is explained further in later chapters.

Even without understanding a single column in this table, or even having heard of any of these banks, it would be fairly easy to assess their performance. Numerical scores on an indicator make it easy to rank banks from high to low. In this case, banks with high capital ratios did well, banks with low capital ratios faltered. Even without knowing what these percentages stand for, or how they've been calculated, a layman can get a basic overview of how banks in Europe are performing vis-à-vis each other, and in general.

THE MULTIFACETED REGULATORY POWER OF INDICATORS

What makes performance indicators like this so popular is precisely this, their ability to reduce complexity and allow information about organisations to be processed easily (Pollitt, 2018; Porter, 2015). Indicators typically present themselves as a product of science, answering the call for more evidence-based policy; they allow policy makers to make decisions based on clear quantitative performance outcomes (Davis, Kingsbury, & Merry, 2012a; Van Dooren, Bouckaert, & Halligan, 2015). Indicators can be used to steer behaviour by increasing learning opportunities, triggering a change in policy strategy, holding underperformers accountable, fostering competition between organisations, or allocating scarce resources (Braithwaite, 2014; Kagan, 1995; Levi-Faur, 2005; Rottenburg & Merry, 2015). In the case of the stress test, banks who do not do so well are required to raise capital or de-risk their portfolio.

Performance indicators are omnipresent in every layer of governance, across many policy domains. Take for instance educational policy. Schools do not have to go through a 'stress test', but governments rely on PISA, the OECD's 'Programme for International Student Assessment'. PISA evaluates educational systems by measuring 15-year-old students' scholastic performance on mathematics, science and reading (OECD, 2019). Approximately seventy countries participate, and the results have been displayed in figure 1 below (FactMaps, 2016). Again, this numerical, and colour coded, representation makes it easy for anyone to see who is performing well (above 500), and where improvement is needed (below 450).

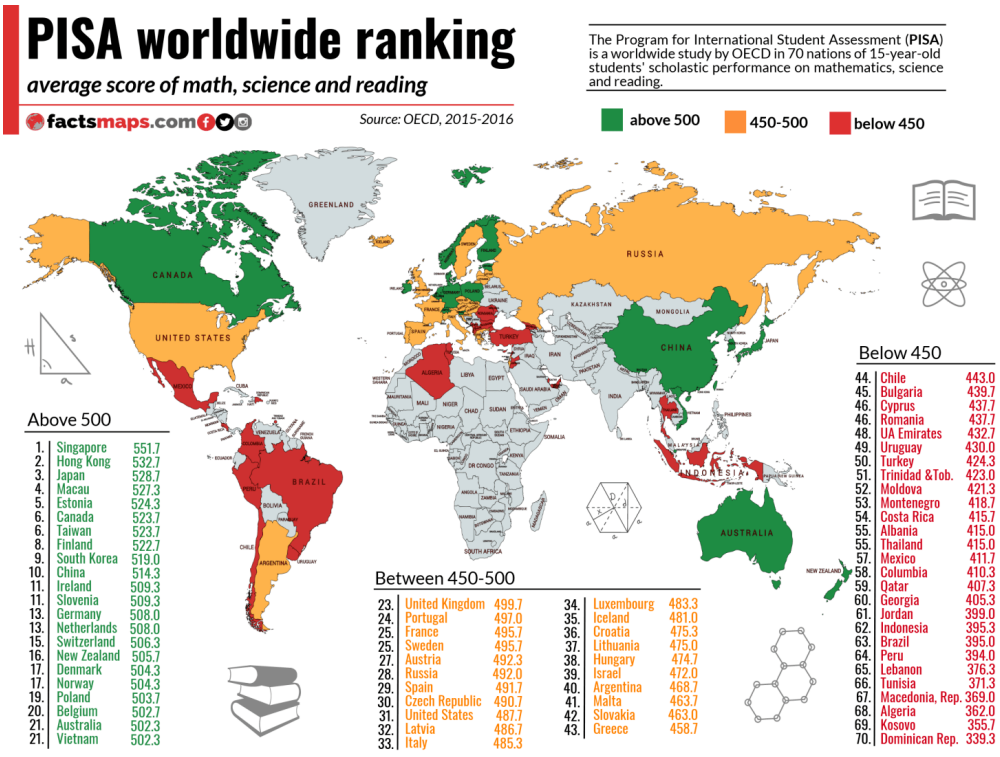

FIGURE 1: PISA WORLDWIDE RANKING

Indicators are important tools for regulation. Regulation is an ambiguous concept that can be defined in various ways. It can be understood broadly as all means used to influence the behaviour of regulated actors, or narrowly as a specified set of formal legal commands. Regardless the definition, regulation always involves three key aspects: information gathering; standard-setting; and behaviour modification (Hood et al. 2001; Lodge & Wegrich 2012; Roness et al. 2008). Indicators fulfil all three conditions, proving of great interest when they are used as regulatory tools in governance, to set standards, improve performance, and (in)formally demand compliance.

Performance indicators are sometimes formally anchored in hard law. For instance, since 2014, the results of the stress test feed into banks' Supervisory Review and Evaluation Process (SREP), which amongst others determines how much capital banks are required to hold (EBA, 2019b). However, this is not always the case. Indicators often exert considerable soft power (such as peer pressure or reputational damage), even without any legal backing (Boswell 2008; Davis et al. 2012b; Espeland 2015). For instance, the OECD's PISA has no formal legal consequence for national educational policy (Gorur, 2016). However, a report of the policy impact of PISA found "over 85 percent of policy makers, local government officials, academics and researchers report having a relatively high level of knowledge of PISA processes and impact" (Hopkins, Pennock, Ritzen, Ahtaridou, & Zimmer, 2008, p. 19). Indicators and rankings are also readily picked up by news media, which makes it difficult for policy makers to completely ignore the results. Thus, whether legally imbedded or not, the regulatory impact of

performance indicators should not be underestimated. Especially in transnational contexts, indicators are convenient tools to steer behaviour in policy areas where supranational institutions have few formal competencies (Scott & Trubek, 2002). This is discussed in more depth in the next chapter.

In addition to this, indicators can steer behaviour in other, and perhaps more encompassing, ways: through the measurement process. Processes of quantification, or making things measurable, have far-reaching, and often unrecognised, implications (Bijker & Law 1992; Callon & Muniesa 2005; Espeland & Stevens 1998; Miller 2004). This is often overlooked in current literature. In this dissertation I am especially concerned with how the measurement process steers the behaviour of regulators and regulated entities. The overarching research question is: **What are the regulatory implications of the social processes through which indicators are calculated?** I answer this question in three parts: I look at the regulatory implications of how indicators measure, manage, and make through processes of calculation.

BREAKING THE BANK?

The aim of this dissertation is to gain a fine-grained understanding of the regulatory implications of knowledge production through indicators. To do so, I selected a single case to study in-depth: The EU-wide banking stress test, briefly referred to above. I am interested in how producing and calculating the stress test affects regulators, regulated entities, and the relationship between them. I opted for the banking stress test for several reasons. First of all, finance as a policy domain combines strong technocratic

institutions and established professions, with high political stakes and strong political agendas. This makes it a good case to consider both political and technocratic drivers in knowledge production. Secondly, it is a recent indicator, still in full development. This makes it especially interesting to look at how different iterations of the stress test have differing impacts on behaviour over time. Moreover, this makes it more practical to speak with respondents involved in the indicator's design. Finally, little socio-political research has been done regarding performance indicators in the financial sphere (an important exception is Farlow (2015)), despite their societal salience; especially in the aftermath of the 2008 crisis and accompanying public bail-outs.

My research question is concerned with how indicators regulate. When applied to the case of the stress test, it indirectly deals with the question whether the stress test is a good regulatory tool, i.e. is it **'breaking the bank?'**. As such, each empirical chapter addresses both the general research question, and this case-specific one. 'Breaking the bank' can be interpreted in three distinct ways. Firstly, the saying typically means that something costs more than one can afford. A reoccurring criticism regarding indicators (and this is certainly the case for the stress test) is whether their cost is worth the benefits: do they actually make a difference, or are they just a waste of resources? (Hood & Peters 2004; Johnston 2004; Pollitt & Talbot 2004; van Thiel & Leeuw 2002). Secondly, the question whether the stress test is breaking the bank, can be understood as a question of whether the stress tests is tough enough on banks. This has to do with a common criticism that indicators are weakened by industry interests 'capturing' regulators

(Baxter, 2011; Hanegraaff, Beyers, & De Bruycker, 2016; Klüver, 2013). Finally, the question can be interpreted as a question of whether the stress test can help us understand which banks are broken, and why. It is commonly said that indicators often act as a black-box, making it unclear where results come from (Enria, 2018).

SO, WHAT?

Indicators steer the behaviour of organisations that affect our daily lives. We rely on indicators of health, environment, employment, infrastructure, and finance to tell us how our societies are doing, and regulate the behaviour of the organisations operating in them. Indicators have an impact on the curricula of schools; shaping how children in society will be raised. They have an impact on how our health care systems are run, and how our financial institutions operate (Boswell 2008; Davis et al. 2012b; Kaufmann & Kraay 2007; Kelley & Simmons 2015; Mol & De Kruijf 2004). Despite indicators rapidly multiplying as tools of (global) governance (Davis et al., 2012b; Rottenburg & Merry, 2015), their wide-spread use has not been accompanied by much systematic reflection on how they are made.

Especially in transnational settings, the complexity of cross-border policy issues simultaneously enhances both the need for clear-cut indicators, as well as the difficulty to create such indicators across varying contexts. In contrast to their air of objectivity, it is everything but self-evident how these indicators are defined and measured (Davis et al., 2012b). As such, before

we can use these numbers in policy and public debate, we need a thorough understanding of where they come from, and what they do.

OUTLINE OF THE DISSERTATION

The question this dissertation addresses is: **What are the regulatory implications of the social processes through which indicators are calculated?** I take an interpretive approach to this study, focusing on how actors involved in the production of the indicator make sense of this process and how they are affected by it.

The next chapter embeds my research in the wider literature on performance measurement and processes of quantification. I link the development of indicators to wider societal trends of modernity and state-control (Scott, 1998). The use of indicators can be traced to the seventeenth century (and probably even beyond that), as statistical methods and nation states developed in symbiosis (Desrosières, 1993; Porter, 1995). I describe how New Public Management (NPM) ideas of efficiency and performance measurement, replaced bureaucratic traditions, causing an explosion of quantitative performance information in what I call a 'hyper modern' society (McLaughlin, Osborne, & Ferlie, 2002).

The third chapter provides an overview of the methodologies and methods I used. I clarify what I did, how I did it, and why. I start with the 'why', providing a brief overview of methodological principles that guide my research practice. I specifically elaborate on my interpretive stance and my use of Actor-Network-Theory (ANT). This particular approach allowed for a

new way of looking at the process of how indicators are calculated. Rather than interpreting calculation as a mechanical process of finding an objective way to represent an essentialist truth about the world, my research heuristic challenges these assumptions, and in doing so allows for fresh insights on the regulatory implications of this process of knowledge production. Moreover, by combining ANT and interpretive research, I enrich the latter by not only focussing on meaning-making practices of human actors, but additionally on the agency of non-human actors in creating meaning. I also elaborate on 'what' I did and 'how' I did it; my methods of data generation and analysis. In conclusion I summarize what my data can and cannot do, to ensure a better understanding of the following empirical chapters.

The next section of this dissertation are the three empirical chapters that all deal with the question of the regulatory implications of how indicators are calculated in the case of the EU-wide banking stress test. They discuss how indicators measure, manage, and make. These chapters also all address the question whether these stress tests are 'breaking the bank', each from the distinct points of view addressed above and repeated here.

The first empirical chapter, chapter four, deals with how indicators measure. A key mechanism in measuring an indicator, is making the various entities comparable according to a common metric, this is called commensuration. Despite a substantive literature, little empirical work has been done to further our understanding of the social and political processes that drive this commensuration process, and its regulatory implications (Peeters & Verschraegen, 2013). I contribute to existing literature by explaining why, despite a preference for commensuration, regulators allow

incommensurability to exist. The question of whether stress tests are breaking the bank, can be interpreted as the question of whether stress tests are tough enough. I show how measurement 'bias' is used to regulators advantage in the stress test.

Chapter five deals with how indicators manage. This chapter borrows from Merton's functional sociology (Merton, 1968) to further the understanding of the disfunctions and latent functions of indicators. My contributions in this chapter lie in explaining how certain disfunctions do not necessarily hamper performance outcomes, and how certain latent functions contribute importantly to performance outcomes. I address how measurement processes, rather than solely measurement outcomes, can steer behaviour. The question of whether the stress test is breaking the bank is dealt with here as whether the costs of the stress test are too high. I argue that in order to weigh the costs and benefits, we need to include the latent regulatory benefits of the exercise.

Chapter six deals with how indicators make. I argue that indicators do more than simply summing up pre-existing characteristics of the world. Rather, they actively (re)make the way we see the world, and ultimately what we accept as truth. In this chapter I explore how using indicators based on large data sets is accompanied by a 'Big Data State of Mind', the epistemological notion that one can or should rely on large data sets rather than theory to observe and understand reality. I discuss how this affects regulation, and calls in question traditional notions of accountability and transparency. I contribute to a new interpretation of transparency, focused on dialogue rather than disclosure. This chapter addresses the question of

whether the stress test is breaking the bank, as a question of whether the stress test succeeds in telling us why banks 'break'. I argue that the stress test often operates as a black box, calling for new accountability mechanisms.

The concluding seventh chapter formulates an answer to the research question 'what are the regulatory implications of knowledge production through indicators?'. Overall, I argue that the process of producing indicators is regulatory in itself, i.e. it steers behaviour. I elaborate on how indicators measure, manage and make, arguing that indicators steer behaviour through largely unrecognized processes of commensuration, governmentality, and epistemic shifts. I also provide an in-depth discussion on whether the stress test is 'breaking the bank'. More broadly, I consider the wider implications of my findings and provide recommendations and avenues for further research.

THE TAKE HOME MESSAGE: CONCLUSIONS ON THE STRESS TEST AND BEYOND

The objective of the EU-wide banking stress test was to "assess the resilience of financial institutions to adverse market developments, as well as to contribute to the overall assessment of systemic risk in the EU financial system" (EBA, 2019a). Did it succeed in doing so? In the following chapters I tackle this question in more detail from various angles. For the curious reader, I already provide a brief summary of my findings.

Overall, I would argue that the stress test has made important contributions to assessing the resilience of the financial system. Where the exercise was widely accused of being biased and not severe enough to truly

stress banks (Dowd, 2015), I would argue that regulators were well aware of the performative effects of public assessments of health and risk, and were mindful of the risk assessments they could 'afford' in a context of government austerity and a lack of EU-wide backstops. Although the stress test might have seemed mild in its results, in its execution, it was a very severe exercise for banks. It forced banks to professionalise their internal risk management; developing new IT systems, digging deep into their own data, and improving communication across risk departments. Beyond allowing supervisors to assess the resilience of institutions, the institutions themselves are now much better equipped to assess their own risk and face future crises.

A worrying trend is the complexity and opaqueness of the top-down quality assessment and the methodological constraints. A first downside is that they reduce banks incentives to invest in their internal risk management, because their results are eventually overridden. A second, and more important, downside is that it becomes difficult to assess where banks' results come from, and if these knowledge claims are valid. I suggest that a better dialogue between banks and regulators would contribute importantly to the EBA's objective of assessing the resilience of financial institutions and systemic risk.

More broadly, what this dissertation will show is that the process of calculating indicators entails important regulatory implications. It is not just the results of indicators that can be used to steer behaviour, rather, the measurement process is steering in itself. Practitioners should not only consider how an indicator can be designed to accurately and objectively measure performance. Instead, practitioners should consider how the

measurement process affects performance outcomes in itself. Failing to consider the regulatory effects of the measurement process, can and will eventually hamper performance outcomes. The regulatory effects of knowledge production through indicators have been largely overlooked in literature, and warrant more scholarly attention. Knowledge production is not simply a matter of measuring organisations' behaviour; knowledge production fundamentally regulates the behaviour of organisations that affect our lives on a daily basis.

WHAT WE KNOW AND WHAT'S LEFT TO LEARN

Chapter overview

Performance measurement in the public sector has a long lineage. In this chapter I provide a brief overview of the history of performance measurement and performance indicators in the public sector. The origins of performance measurement can be traced to the seventeenth century (and even beyond), as statistical methods and nation states developed in symbiosis. In the 1980's New Public Management (NPM) ideas of efficiency and performance measurement caused an explosion of quantitative performance information, that has not been exempted of critique. Performance indicators play an important role in the way governments distribute attention, make decisions, and allocate scarce resources. At the end of this chapter, I argue that despite their widespread use, we know little about the social processes through which performance indicators are made, and to what effect. Understanding how performance indicators are made is increasingly important as proliferating technologies of quantification have far-reaching and often unrecognized governance implications.

THE SIMULTANEOUS DEVELOPMENT OF STATISTICS AND THE STATE

MEASUREMENT AS A PROJECT OF MODERNITY

> "Now, what I want is, Facts. Teach these boys and girls nothing but Facts. Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only form the minds of reasoning animals upon Facts: nothing else will ever be of any service to them. This is the principle on which I bring up my own children, and this is the principle on which I bring up these children. Stick to Facts, sir!"

These are the opening lines of Charles Dickens' novel Hard Times (1854, p. 3), pertinently capturing the Zeitgeist of modernity. Once upon a time, Max Weber (1919) wrote, we lived in a world driven by mysterious powers of nature, spirits and Gods. But, this magical world became 'disenchanted' through processes of rationalisation[3]; the world became 'modern'. Science and scientifically oriented technology were the motive force behind these processes of modernity; promulgating the belief that one can learn anything at any time. As Weber continues (1919, p. 8): "there are no mysterious incalculable forces that come into play, one can, in principle, master all things by calculation".

[3] Weber acknowledges that processes of rationalization predate modern times. The ancient Hebrew prophets already eschewed magic. However, Weber states that the urge to disenchant reaches its logical conclusion in the seventeenth-century (Bennett, 2011).

Weber, but also Michel Foucault (1977) and Anthony Giddens (1990) note that modernity not only gave way to scientific and technologic progress, but also to secularism, individualism, and the bureaucratic nation state. The belief arose that societies can, like the rest of the world, be understood, measured, and improved. In the remainder of this first section, I provide a historical overview of the modern pursuit of measurement and control and how it ties in with the development of the bureaucratic nation state.

It is important to mention that some authors, such as Lyotard (1984) and Baudrillard (1994), believe that modernity ended in the mid- or late 20th century and we now live in an age that can be marked 'Postmodernity'. Other theorists, however, suggest that our society is in a 'liquid' (Bauman, 1989), or 'high' (Giddens & Pierson, 1998) modern phase, rather than a postmodern one. They argue that key features of modernity are still present in current day society, albeit in a different way than the early modern period. Where early modernity generated scepticism towards traditional narratives (religion, tradition, superstition, dogma's), current day modernity generates scepticism towards the master narratives of modernity itself. Reflexivity and continuous questioning have always been central to modernity, right from the time of the Enlightenment (Giddens, 1990). High modernity is characterised by radical reflexivity, the continuous monitoring and revision of beliefs in the light of changing evidence or circumstance. This only perpetuates the need to measure, measure again, and measure more[4].

[4] Latour (1991) adds to this debate an interesting critique that 'we have never been modern'; that as a society the distinctions we believe to have drawn between facts and values, nature and society, superstitions and science, humans and things, simply never held (nor will they ever).

MAKING SOCIETY LEGIBLE

The evolution of science and modern statistics is inextricably connected to governance and state-building. As the etymology of the word already shows, statistics is connected with the construction of the state, with its description, unification and administration. In his seminal book 'The politics of large numbers', Alain Desrosières (1993) provides a detailed account of the history of statistics. From the seventeenth century onward, states began to record marriages, births and deaths in central registers. Keeping these records stemmed from a modern administrative and judicial concern of fixing (and arguably also creating) the identities of individuals and their links of kinship within a nation state. In the eighteenth century, more advanced statistical techniques made it possible to use this registry to calculate population trends. Social questions – no longer random events or the will of God, but statistical patterns - became a concern of the state. The patterns of crimes and suicides were no longer attributed to individual agency, but were seen as properties of "society"[5], and, from the nineteenth century onwards, they were widely considered to be the best evidence for its real existence (Polanyi & Nye, 1962; Porter, 1995).

Statistics were essential for statecraft. This idea is explored in Scott's (1998) book on (failed) projects of modernity. He explains how the illegibility of localised and diverging measurement practices was an administrative headache for monarchies, who found it impossible to control their vast

[5] See for instance Durkheim's (1897) ground-breaking book 'Suicide', the first to show that suicide had social causes rather than just being a matter of individual temperament.

territories, sometimes including oversea areas. Scott (1998, p. 55) quotes Benjamin Constant stating:

> "The conquerors of our days want their empire to possess a unified surface over which the superb eye of power can wander without encountering any inequality which hurts or limits its view. The same code of law, the same measures, the same rules, and if we could gradually get there, the same language; that is what is proclaimed as the perfection of the social organisation… the great slogan of the day is uniformity."

Converting units of one region to another was at least an inconvenience, if not an obstacle, to state expansion and the growth of large-scale trading networks. Take for instance the French Revolution in 1789, where a key to its success was standardizing the territory. Precise, uniform measures enhanced administrative control over matters of taxation and economic development (Desrosières, 1993; Porter, 1995).

So too, the success of colonial empires hinged on measurement techniques to control oversea territories and regulate the daily lives of locals. For instance, Mitchell (1988) tellingly describes the colonisation of Egypt as the establishment of a principle of order. In order to subjugate the local population, their place was carefully marked out, their quota were specified, and their performance continuously measured. Regional and central 'Bureaux of Inspection' were established to monitor the daily performance records. Mitchell (1988, p. 33) writes: "Egypt was to be made picture-like and legible, rendered available to political and economic calculation. Colonial power required the country to become readable." In the same vein, what are

labelled 'Neo-Colonial' efforts are furthered through measurement techniques too. De Maria (2008) writes how the Western mission to measure and control corruption in poor African countries is conducted on the terms and in the style of the West. The result is a rapidly growing, yet arguably failing, anti-corruption movement, detached from street level realities.

REPRESENTATION AND STANDARDISATION

Measurement and representation is necessarily transformation (Latour, 2013). People, places and things have to be remade before they can be represented, mapped, or counted in distinct categories and classifications. There is much of what Weber called rationalization in this, and also a good deal of centralization. Geography is an important discipline in this respect as it long appeared to unproblematically measure and map the world, paying homage to the illusion of one-to-one representation (Lindner, 2017).

Establishing classifications is a question of taxonomy; examining the nature of the bonds that hold people and places together. Importantly, classification and taxonomy are a matter of power, the power to define and structure reality (Stone 2012). For most of the human sciences, problems of classification and coding are often perceived as mere technical and practical problems, to be solved from one day to the next by practitioners, rather than by theoreticians. However, some subdisciplines of human science are entirely attributed to studying how these classifications come to exist and what their implications are, most notably studies on the history, philosophy, or sociology of science (by thinkers such as Foucault, Gadamer, Kuhn, Popper,

Gillespie…). Here classifications are not merely discussed as grids or tools through which the world could be discussed, but they are considered objects of study in themselves[6]. I will revisit this body of literature in more detail later on in this chapter. What is important, is that according to this perspective these classifications do not merely describe an independently existing world around us, but they actively shape the world they attempt to depict.

Statistics of probability helped state officials master the unpredictability of social life (at least to a certain extent). This venture was explored by insurers using mortality tables in the eighteenth century, but it was only fully championed thanks to Belgian astronomer Adolphe Quetelet in the mid nineteenth century, as he pioneered the regularity of averages, and the law of large numbers, to structure and predict individual behaviour (Desrosières, 1993). Quetelet reconciled individual diversity with statistical regularities along a Gaussian curve, as a law of errors, distributing particular individuals around his famous (and idealized) 'average man'. This mode of thinking centred on averages and standard deviations overcame and structured the chaos of countless singular observations that were perceived as random and presumed unmanageable or unavoidable. Thanks to the laws of probability, states could make rational choices in situations of uncertainty; as well as evaluate individuals according to their position on Gaussian curves and deviations from the standard[7].

[6] Take the statistical label 'Hispanic Americans', that lumps together Americans of Mexican, Cuban, Puerto Rican, Iberian, as well as Central and South American decent; who do not by any means universally support this label (Gillispie, 1960).

[7] Authors such as Michel Foucault, Ian Hacking and Nikolas Rose have been especially profuse in criticising the authority of statistical norms through which a language of (ab)normality is created, stigmatizing those who fail to conform (Schaffer, 1989).

BOOMING BUREAUCRACIES

This culture of quantification and control only grew. Governments held an unwavering 'trust in numbers'; which Theodore Porter (1995) unpacked in his thusly named book. His first sentence beguilingly reads: ""OBJECTIVITY" arouses the passions as few other words can" (emphasis in original) (Porter, 1995: 3). This was in line with the thinking of the Progressive Movement, led by president Woodrow Wilson, often considered the father of US public administration. The Progressive Movement argued that states needed an administration to "discover, first, what government can properly and successfully do, and, secondly, how it can do these proper things with the utmost efficiency." (Rabin & Bowman, 1984, p. 136). Objective bureaucratic administrations were cut loose from politics, and relied heavily on scientific knowledge (Barzelay & Armajani, 1992). This ensured a consistent application of universal rules that benefited the public interest, undermining the power of political (and biased) party machines.

With the popularity of standardisation rising in the post-war period, criticism of this culture of objectivity also resurfaced[8]. Most notably from academics in the tradition of the Frankfurt School, such as Adorno and Horkheimer in their Dialectic of Enlightenment (1947). They accused the quantification of the culture industry for the hollowness of mass culture, blaming calculation and statistics for replacing unique and curious individuals with boring and meaningless median citizens, stripping them of all true meaning and value. Statisticians boasted that their science averaged away

[8] In the nineteenth century faith in numbers was already commonly ridiculed. For instance, to put the sexes in balance, it was proposed to marry one and one half man with three women minus a quarter per kilometre squared (Belpaire, 1847).

everything random, accidental, or inexplicable, and left only large-scale predictable patterns and laws. The Frankfurt School condemned them for it. This to show, quantification has the vices of its virtues. Another important group of critics were actuaries and accountants themselves, of which some preferred to use their expert, albeit more subjective, interpretations, rather than some mechanical standardized calculation (Bryer, 2000).

BUREAUCRACY ISN'T ENOUGH

NEW PUBLIC MANAGEMENT

By the end of the twentieth century, these stable bureaucracies proved too slow, burdensome, impersonal and unresponsive to a fast-changing, globalising society. Buried in paper work and red-tape, they often lost track of the outcomes of governments; what was being done. These undesirable side-effects of bureaucracies interfered with effective governance (Barzelay & Armajani, 1992; Drucker, 1968; G. J. Miller, 1992; Osborne & Gaebler, 1992; Zifcak, 1994). Moreover, by the late twentieth century, the needs of citizens had changed. Rather than 'basic' one-size-fits-all services, citizens required tailor-made solutions designed for their individual needs; and wanted to have a say in the design and delivery of these services (McLaughlin et al., 2002).

In order to continue the modernist objective of measuring and controlling society, a reformation was needed. Governments (re)turned[9] to

[9] Bureaucratic-reforms were also partially inspired by business , most notably the scientific management paradigm that was established to identify the most efficient rules and procedures (Barzelay & Armajani, 1992).

the private sector for inspiration. Corporations had already begun restructuring to adapt to the ever-changing, complex society. Management reform led to a more customer-centred, decentralized approach, with clear performance targets and incentives to encourage competition (Barzelay & Armajani, 1992). Governments sought to mirror these institutional and organisational aspects, to become more customer-driven and service-oriented. This government reform was heralded with books titled 'Breaking Through Bureaucracy' (Barzelay & Armajani, 1992), 'Reinventing Government' (Osborne & Gaebler, 1992), and 'Managerial Dilemmas' (G. J. Miller, 1992). All of them imploring governments to deal with the undesirable consequences of their bureaucratic administrations.

A new paradigm emerged, lumping together these ideas for a more entrepreneurial government that was run like a business: New Public Management (NPM). NPM aggregates a plethora of policy principles that all in some way intend to improve public sector performance by making it more efficient and goal-centred (Van Dooren et al., 2015). Different authors have emphasised different aspects of NPM, but the main overlapping principles are disaggregation and decentralisation, competition, and incentivization (Dunleavy, Margetts, Bastow, & Tinkler, 2005). Christopher Hood's (1991) article 'a public management for all seasons' provides a comprehensive overviews of the origins, rise, and conception of NPM. Hood argues that NPM presented itself as a universally applicable framework, being both politically neutral, as well as portable across policy fields and national contexts. Most notably, public sector performance measurement and management systems were introduced as governments pledged to 'do more with less' (Gore, 1993).

PERFORMANCE INDICATORS & REGULATION

With NPM, the measurement and use of performance information became more institutionalised and more professional (Desrosières, 2015; Van Dooren et al., 2015). Performance indicators proliferated as fiscal stress and legitimacy crises pressured the public budget and politico-administrative system. Governments and their agencies were graded, red-lighted, and ranked in an effort to provide insight into the muddy waters of efficient and effective government (Ingraham, 2005; Van Dooren et al., 2015). Especially the Anglo-Saxon world witnessed a massive boost in performance indicators by the end of the 1980s. Though not with the same intensity, performance information also became pivotal in public sector reform in continental Europe (Van Dooren et al., 2015). Macro-economic indicators, service quality measures, client surveys, opinion polls, audit reports, Key Performance Indicators, and programme evaluations bourgeoned. The simplicity of these performance indicators allows for more succinct communication with actors inside and outside government, tapping into an agenda of transparency and accountability; at least in theory (Sarfaty, 2011).

The rise of NPM and performance indicators uncoincidentally coincides with a shift towards a 'regulatory state' (Majone, 1994). Indicators serve the main three objectives of regulation: information-gathering, standard-setting, and behaviour-modification (Hood et al. 2001; Lodge & Wegrich 2012). The 'regulatory' state replaces the traditional sovereign state model, with its command-and-control policy style, public ownership, and nationalisation (Christensen & Lægreid, 2006). Instead, the regulatory state comes with a rise in privatisation and liberalisation, with the objective to

improve efficiency and promote competition while protecting consumers and citizens. The rise in autonomous formal organisations created a need for more formal and objective regulation (Brunsson & Sahlin-Andersson, 2000). Moreover, as policy issues became increasingly complex (with the rise of integrated global economies, free movement of goods and services, and environmental concerns), trans- and supranational regulation gained in significance, complementing the regulatory capacities of nation states (Majone, 1994).

Indicators play an especially important role in transnational regulation. Besides indicators anchored in formal law, informal indicators can be used to obtain steering power in policy areas that are (mainly) the formal responsibility of national governments (Scott & Trubek, 2002). Take for instance the European Union's Open Method of Coordination (OMC). The OMC sets broad, non-binding, policy goals and Member States agree on indicators to measure best practices. Although the policy goals are non-binding, through naming and shaming the OMC attempts to steer Member State's behaviour covertly (De Ruiter, 2008)[10]. According to Héritier (2017) this kind of 'covert' integration of policy making, is politically more expedient and less costly. Pollack (2003) calls this integration 'by stealth'. However, De Bièvre & Bursens (2017) warn that covert integration measures can also tactically be used to shy away from further (overt) integration through legislation, and leave policy spaces vague. Moreover, because covert integration takes place outside the formal European political decision-making arena (at least to some extent), Héritier (2017) notes that it also creates

[10] Incidentally, the OMC is yet to lead to any significant integration (Schäfer, 2006)

problems of democratic legitimacy[11]. This ties in with concerns voiced by Davis et al. (2012b) that transnational indicators promulgated by extra-national entities, give these (undemocratic) entities the power to steer the behaviour of organisations that affect our everyday lives. They provide the example of the World Bank that prompted many countries to reform their legal systems simply by publishing their country-level indicators on the ease of doing business.

It should be noted that, beyond the NPM paradigm, the last decades have witnessed other reform initiatives as well; most notably the Neo-Weberian State (NWS), and New Public Governance (NPG). Where NPM heavily emphasises efficiency, the latter two reform paradigms aim to strengthen the effectiveness and legitimacy of government (Junjan, 2015). Despite the slight differences in emphasis, performance indicators remain pervasive tools across these paradigms.

FROM KNOWLEDGE USE TO KNOWLEDGE CO-PRODUCTION

CONTESTING CALCULATION

The use of performance indicators has not gone uncriticised. A first strand of critique comes from public administration scholars, exposing adverse effects of performance indicators (Pollitt, 2018). They are said to lack accuracy, encourage gaming, demotivate workers, be biased towards what is

[11] Although this is a valid concern, De Bièvre & Bursens (2017) note that the problem might not be so severe in some cases, where for instance national actors take over from executives. Moreover, when covert integration is used to overcome political deadlock, it might demonstrate high levels of output legitimacy by delivering acceptable output for citizens.

quantifiable, and ultimately not substantially improve performances (Berten & Leisering, 2017; Bevan & Hood, 2006; Bouckaert & Balk, 1991; Davis et al., 2012b; Hvidman & Andersen, 2014). Recent research has also identified misuse of performance information as a key factor hampering performance improvement (Micheli & Pavlov, 2017; Moynihan & Kroll, 2016; Taylor, 2011; Van Dooren et al., 2015). Moreover, on the practitioner side, not all public officials are happy to see their work reduced to a number on a scale. They argue that performance indicators do not do their work justice, overlooking important aspects of their job that are not so easily quantified. Performance indicators cannot capture front-line complexity, ignoring important local knowledge (Durose, 2009). The example of the Soviet Union's, failed, planned economy is often brought forward as a key illustration of how measuring targets puts quantity over quality, and hampers rather than improves actual performance (Bevan & Hood, 2006; de Bruijn, 2001).

This feeds into a second criticism drawing on critical theorists such as Adorno and Horkheimer (1947), Bauman (1989) and Habermas (1984), mentioned earlier. Contemporary Critical authors (e.g. Denhardt, 1981; Dunn & Fozouni, 1976; Dunn & Miller, 2007; Fischer & Forester, 1993) argue that NPM and systems of performance measurement are unable to go beyond instrumental rationality or incorporate forms of hermeneutic and critical reason (a critique I partially refute in chapter four of this dissertation). The instrumental rationality present in the NPM discourse is said to foster a decline of democracy and individual freedom (Mouzelis, 1967).

A third criticism comes from the field of Science and Technology Studies (STS). STS scholars warn for an overreliance on evidence-based policy

and technocratic decision making as a whole. Some go so far as to claim evidence-based and rational policy making is a myth, because policy work is fundamentally political (Boswell 2018). STS scholars argue to dismiss the idea of a 'double-delegation', as Callon, Lascoumes, and Barthe (2009) call it, where knowledge about the natural world or society is seen as something external to policy; as a separate entity produced by experts in research centres and then (mis)used by politicians. This provides science, and performance indicators, with an aura of neutrality and objectivity. As if performance indicators were produced in a vacuum, free from any political consideration. Rather than criticizing politicians for cherry picking in scientific reports, this group of scholars argues that scientific facts and findings do not exist as some distinct, objective entity, to then be politicized. Instead, knowledge is socio-political from the off-set. The design and production of performance indicators is contingent to the socio-political context (Jasanoff, 2006). Deborah Stone (2002) adds that numbers work a lot like metaphors; they select one feature, and ignore others. Counting requires judgment about inclusion and exclusion. Every performance indicator is a political claim about where to draw the line.

THE COPRODUCTION OF KNOWLEDGE AND POLITICS

Knowledge and politics are produced in simultaneous processes. Measurement means codification, codification means choice, and choices are political. This touches on the notion of knowledge 'coproduction', the idea that knowledge and politics are produced in simultaneous processes. Performance indicators are made by networks of actors and institutions, with

limitations, expertise and interests. Science is influenced by cultural and political institutions, which are in turn dependent on science. As Latour states (1983: 168) "science is politics by other means". This means that although modernity calls for a rational and purely scientific understanding of the world, there is no such thing. More scientists should then be concerned with not only studying the outside world, but with studying themselves, studying how scientific knowledge is made, and to what effect. This is what Bloor (1981) called the principle of symmetry, the idea that all knowledge (both myths and science) should be treated 'symmetrically', as equally in need of explanation, as they are equally socially constructed. This principle is a founding tenet of the STS field. Science nor politics holds the privilege of producing 'the truth', truth is negotiated (Latour, 2013). Another important work in the field is Thomas Kuhn's 'Structure of Scientific Revolutions' (1962)[12]. This was one of the first books to acknowledge the paradigmatic evolution of scientific knowledge, explained better by shared societal psychologies than resemblance to a so-called external and objective reality.

A good example of knowledge coproduction is found in the development of the Gross Domestic Product (GDP), one of the most well-known performance indicators that assesses how well our economies are performing. In his popular book 'Gross Domestic Problem', Lorenzo Fioramonti (2013) uncovers the political influence in "the world's most powerful number". GDP was originally designed in the 1930's to help America come out of the Great Depression. Roosevelt's New Deal policy rested on the assumption that the state should measure the efficacy of its

[12] Although see Stephen Turner's (2008) instructive chapter on the social study of science before Kuhn.

policies and intervene where necessary. And so, the GDP was developed to help the government allocate funds to the most efficient policies. After the war, the United States and United Kingdom took the lead in standardizing this measurement throughout the United Nations[13]. In this standardisation process the choice of how to measure if the economy was performing well, which parameters to include and how, was a practical and political choice, rather than a fully 'rational' or 'scientific' one. It was simply convenient to value commercial goods at market price, government services at cost and completely ignore unpaid household activities or ecological costs[14]. Over time, the random imaginary line dividing productive and unproductive activity, became a real-life 'fact'. Worldwide, all kinds of policy measures are implemented and defended in the name of fostering GDP-growth, randomly benefiting policy regarding industrial production but undervaluing policy measures that involve technological innovation for instance. GDP thus represents a certain model of society, influencing government policies and priorities. This example of GDP, illustrates the notion of knowledge coproduction well: politics produce knowledge (the politically driven pragmatic design of GDP), and knowledge produces politics (GDP affects policy funding and development).

[13] Critics also condemned the complicated statistical construction of the GDP, complaining that very few people actually understand how the figures are produced, and asking what such a complex abstraction can actually mean (Fioramonti, 2013).

[14] Not wholly unsurprisingly, in 2009, the French government asked the prominent economists Amartya Sen, Joseph Stiglitz and Jean-Paul Fitoussi to propose revisions for the calculation of GDP. They suspected that GDP was a poor measure of the 'wealth' generated by a nation within a year (Desrosières, 2015).

How knowledge is made

Following this body of literature, it is imperative to unpack this process of coproduction, to truly understand the regulatory implications of performance indicators. Indicators are often presented as objective facts and all agency involved in their development is stripped away. There are strong interests to obscure the origins of information and classification schemes and to reify them as facts that speak for themselves and do not require further investigation into how they were produced (Davis et al., 2012a). Despite their widespread use in regulation, the literature on how regulatory indicators are made is scarce (Important exceptions include Bartl et al. 2019; Berten 2019; Cook 2017; Davis et al. 2012b; Thedvall 2012). The literature so far has mainly studied the transformative effects of indicators. For example, studies have documented how people change their behaviour in reaction to being evaluated (reactivity) (Espeland & Sauder 2007), how policy processes and policy options alter (Feron, 2013), and how indicators change global power relations and processes of contestation (Davis et al., 2012b; Gorur, 2016). What is missing, is an understanding of how indicators are coproduced, and to what effect (Latour & Woolgar, 1986).

The field of STS has done a great deal of work on understanding how facts are co-produced. Sheila Jasanoff (2004, p. 3) writes that the aim of studying knowledge coproduction should be to "explore how knowledge-making is incorporated into practices of state-making, or of governance more broadly, and, in reverse, how practices of governance influence the making and use of knowledge." She recognizes two ways to study coproduction. On the one hand there is interactional coproduction, which is epistemological in

nature. Literature in this tradition is concerned with how we come to know what we accept as legitimate truth, and how a small group of actors shapes these beliefs in society. On the other hand, there is constitutive coproduction, which is concerned with how facts, or things, are made. It studies how a point of stability is reached in either a knowledge controversy (where eventually a single truth, or fact, emerges), or in a competition of several artifacts or things (where one design or product eventually gains the upper hand). Seminal authors following this line of study are Michel Foucault (genealogy studies) and Bruno Latour (actor-network theory). This dissertation is mainly concerned with this constitutive coproduction. In the next (methodological) chapter I elaborate more on Actor-Network Theory (ANT).

Succinctly put, ANT studies how (social) realities are created through interactions between people and things, that gather together in so-called 'actor-networks' (Latour, 2005). This means studying ANT is also a study of power. It is a study of how certain actors (also non-human) control others; how they impose themselves and their problem-definitions in an interaction, how they define their respective identities, their mutual margins of manoeuvre, and the range of choices which are open to them. It is important to note that keeping an actor-network together is an ongoing process, never a completed accomplishment, and it may also ultimately fail (Callon, 1984). The actor-network needs to continuously maintain a precarious equilibrium, for facts to hold, and stay put, in reality. Facts are not produced once and for all, they need to be maintained by all the actors supporting them. They are

not a thing, they are a process, they are a social construct. Facts evolve through evolving actor-networks, and power-shifts[15].

OUTLINING A RESEARCH AGENDA

This chapter has demonstrated how processes of modernity have led to a co-development between statistics and states. Since the 17th century, states have relied on science to measure societies and standardise them, in order to govern them better. This led to the development of extensive administrations and bureaucracies, concerned with classifying and codifying society (Desrosières, 1993; Porter, 1995; Scott, 1998). This modernist ideal of goal-oriented measurement reached its pinnacle in the nineteen-eighties, with the ideas and practices of New Public Management (NPM), that were supposed to make government more efficient and adept to complex and globalising societies.

Performance indicators played a key role in NPM, helping governments distribute attention, make decisions, and allocate scarce resources (Hoppe, 2009). Performance information promises to be objective and transparent, thus producing more democratic decision-making through higher levels of accountability. However, over the past decades, NPM and performance measurement systems have received much backlash (e.g.

[15] This preoccupation with power also gave rise to a branch of STS especially concerned with the democratization of knowledge production. The project of understanding the social nature of science was reconciled with a project of promoting socially responsible science (Ravetz, 1971). As such, STS is often at odds with technocratic governance, arguing it creates inequalities and undermines citizens' interests (Latour, 1987).

Boswell, 2018; Espeland, 2015; Gorur, 2015; Kaufmann & Kraay, 2007; Pollitt, 2018; van Thiel & Leeuw, 2002). One key criticism is that these indicators are seemingly created in a vacuum, to then be (mis)used by politics. This chapter showed that indicators are political from the off set, and coproduced by socio-political conditions (Davis et al., 2012b; Jasanoff, 2004).

Understanding how performance indicators are made, and to what effect, is increasingly important as proliferating technologies of quantification affect society in subtle and often unrecognized ways. This dissertation will contribute to our understanding of the regulatory implications of knowledge (co)production through indicators, with a case study of the EU-wide banking stress test.

METHODOLOGY & METHODS

Chapter overview

I will be exploring the EU-wide stress test as a performance indicator for European banks. The research that I present in this thesis is the result of an interpretive inquiry, based predominantly on interview material. Before delving deep into the data I have collected, it is imperative to clarify the nature of the knowledge yielded from these interviews. What can this data do, and what can't it? In the following I hope to shed some light on how exactly I did what I did, and why. I will start by taking a brief detour into my philosophy of science, to explain 'why' I did what I did. Here I briefly describe the ontological and epistemological stance taken and lay out the methodological principles that guide my research practice. Following this, I elaborate on 'how' I did it, my methods of data generation and analysis. In conclusion I summarize the key elements necessary for a better understanding of the following chapters.

WHY? MY PHILOSOPHY OF SCIENCE

> "Believing, with Max Weber, that man is an animal suspended in webs of significance he himself has spun, I take the analysis of [those webs] to be therefore not an experimental science in search of law, but an interpretive one in search of meaning."
> – Clifford Geertz (1973: 5)

I will start with an overview of my research framework, to then position myself a little more precisely in the community of interpretive policy analysis. Finally, I will discuss how my metaphysical positioning informs the standards to which I adhered in gathering and analysing my data.

IN SEARCH OF MEANING

My research falls in the tradition of interpretive policy analysis. My research framework combines particular understandings of the nature of reality, the nature of knowledge, the nature of research and the purpose of research. Drawing on the work of Vivien Burr (1995) and Yvonna Lincoln and Egon Guba (1985) I will succinctly describe the main ideas. These are summarised in table 2 below.

My research begins with the assumption that there is no one objective 'reality', or extra-social point of view. Where realists believe that there are truths, and we must find them, a constructivist ontology understands all truths as socially conditioned and value laden (Gordon, 2009; Schmidt, 2001).

Furthermore, I ascribe to a specific understanding of 'social constructivism', borrowed from Actor-Network-Theory (Latour, 2005) (on which I elaborate below). Succinctly put, in my view of constructivism, 'the social' is constructed by interactions between heterogenous human and non-human actors.

The second assumption is that knowledge does not exist in a state awaiting discovery; we can only know through purposive interaction with the world. As such, phenomena can only be meaningfully understood by an interpretation of the meanings that people assign to them. The miner vs. traveller metaphor, borrowed from Brinkmann and Kvale (1996), clearly illustrates this epistemological stance. A positivist view sees research data as precious metal, it is waiting to be dug up by the miner-researcher. The data is extracted unproblematically, and surfaces unaltered by the mining process. My interpretive framework holds a scepticism towards this idea. Alternatively, the researcher is a traveller. Together with the research subjects s/he explores the field and generates data in conversation. This also implies a reflexive attitude towards the role of the researcher in generating data. As noted in my epigraph, the aim is not to uncover general laws.

My research is thus concerned with understanding how and why people and things make realities, make sense of realities, and the regulatory implications of these processes.

TABLE 2: ELEMENTS OF MY RESEARCH FRAMEWORK

Nature of reality (ontology)	Constructivist: Realities are locally constructed
Nature of knowledge (epistemology)	Interpretivist: Knowledge is co-generated in the research process; constructed by interactions between people and things
Nature of research	Exploring processes of how collectives are gathered together and create meaning
Purpose of research	Understanding rather than explaining

Actor-Network Theory

One can embark on a quest for meaning in a variety of ways: It can be done following hermeneutic-phenomenological, pragmatist, dialectical, discursive, and many more traditions[16]. To be sure, the boarders between the 'different types' of interpretive research can be fuzzy. I still find it useful to position myself to a certain extent and clarify how exactly I will study meaning. In my research endeavour I am interested in how meaning is created in interactions between people and things. In doing so I follow the research heuristic of Actor-Network Theory (ANT) (Callon, 1984; Hackett, 2008; Latour, 2005; Law, 1999). In this dissertation I look at how people and things interact in constructing the EU-wide banking stress test, and how this affects regulation.

[16] Overviews and typologies can be found in e.g. Kvale & Brinkmann (1996), Schwandt (2001), and Wagenaar (2011).

In brief, ANT challenges the essentialist idea that certain things are simply true and others simply false. It opposes the idea that things (even scientific facts) are natural and necessary and that it is possible for knowledge to faithfully represent the natural order (Callon et al., 2009; Latour, 1999). Facts are not seen as inexorable truths waiting to be discovered (or 'mined', as stated above). Rather, facts are constructed to be 'true' or 'real', through interactions between people and (importantly also) things[17] (de Vries 2016). When we observe a part of reality, this is actually an actor-network; it is the product of interactions between people and things. ANT studies these interactions (also called translations[18]) and analyses how and why stability is achieved; how something becomes 'true' or 'real'. This often has to do with power, the power to steer and define these interactions, and how the actor-network is assembled.

For example, in Bruno Latour and Steve Woolgar's (1986) seminal book 'Laboratory Life', the authors conduct ethnographic fieldwork in a scientific laboratory, following the day to day activities of working scientists, in order to examine how a scientific fact, TRF (a molecule), is constructed. Between 1962 and 1968 only a part of the TRF amino-acid chain was accounted for, so a group of scientists created a synthetic replica of TRF to help unravel the rest. However, they still had to prove this synthetic TRF had the same structure as the natural TRF. Whether or not the two compounds were different or identical, was ultimately a matter of (social) construction. It

[17] An article where this is expertly demonstrated is for example Latour's (1988b) account of the pasteurisation of France, or Callon's (1984) article on the scallops of Saint Brieuc Bay.

[18] ANT is also called a sociology of translations (Callon 1984). The idea is that the actor-network is assembled into a fact through four steps or translations. I elaborate on these four translations in the next chapter of this dissertation.

would have been possible to dismiss a difference as minor noise, or to deem it a major discrepancy. Although there was much disagreement between scientists, eventually it was published that the natural and synthetic substances would be accepted as identical[19]. And so, after many interactions between scientists, funders, molecules, amino acids, and lab equipment, TRF became a scientific fact: pGlu-His-Pro-NH2.

When studying a situation (like the one above) we often focus too much on the actions (or agency) of the people involved (Czarniawska, 2014). However, each interaction between people is also shaped and constrained by non-humans, by material conditions. These non-human actors also have agency. Callon (1991) highlights an important distinction between seeing these things as mediators, rather than intermediates. What he means by this is that things are not mere intermediate placeholders that do what is prescribed to them by people (Bijker & Law, 1992). Instead as active mediators, things can actively shape relationships and alter the world around them. A seemingly silly example of this is elaborated in Latour's article on the agency of a door-closer (Latour, 1988a). Rather than hiring an 'unreliable youngster' to fulfil the boring and probably underpaid job of opening and closing a door, it is much more cost-effective to just rely on a combination of hinges, springs, and hydraulic pistons to make sure the door can be opened, and will close itself. However, this replacement (or translation) from porter to hinge is not entirely unproblematic. A door with a powerful spring-mechanism will play the role of a rude porter that slams the door shut. This means that you have to adjust your behaviour in accordance with this hinge,

[19] Although further on in the same paper, the author still toyed with alternative structures of TRF as well (Latour & Woolgar, 1986).

you have to move fast and make sure not to be too close behind anyone else. If futile things, such as hinges, can already prescribe and impose behaviour onto humans, imagine what more complex things can do.

IF YOU JUDGE A FISH BY ITS ABILITY TO CLIMB A TREE

A constructivist positioning inspires a particular set of quality assessment criteria as well. Interpretive research is often held to evidentiary standards that it cannot achieve. These criteria, such as validity, reliability, replicability and objectivity, developed over time out of positivist presuppositions. As such, they stand in stark contrast to the fundamental interpretive understandings my research is built on.

Although an interpretive approach is generally opposed towards fetishism of method and technique (Gadamer, 2004; Law, 2004)[20], this does not justify an 'anything goes' attitude, or a reliance on personal intuition (Mills, 2000). There are certain standards[21] that I adhered to throughout my research. In my study I included such trustworthiness[22] techniques as member checks, thick description of phenomena, reflexivity, and an audit trail so that the process of data generation and analysis would be both visible and

[20] Law (2004) for instance points out that methods do not only help us to understand and describe reality, but help to produce this reality that they understand. He presents a quote from Appelbaum that is worth repeating (Appelbaum, 1995, p. 89, in Law, 2004, p. 11): "My hope is that we can learn to live in a way that is less dependent on the automatic. To live more in and through slow method, or vulnerable method, or quiet method. Multiple method. Modest method. Uncertain method. Diverse method. Such are the senses of method that I hope to see grow in and beyond social science."

[21] The debate on a shared set of standards is still ongoing. The foundations for this can be found in Miles & Huberman's Qualitative Data Analysis (1994) and Lincoln & Guba's Naturalistic Inquiry (1985).

[22] The term trustworthiness, is usually preferred over validity. Though in a sense they are each other's counterparts. Where validity looks for truthfulness in the sense of being representative to a true, external reality, trustworthiness emphasises being true to a deliberate, transparent and ethical research process.

verifiable (Schwartz-Shea & Yanow, 2012; Yanow & Schwartz-Shea, 2006). In the second section of this chapter, where I discuss my data generation and analysis in more detail, I discuss in more detail how I implemented these quality standards.

This also has consequences for the nature of any claims to causality or generalizability. First of all, causality is understood as 'constitutive causality'[23]. Here, human meaning making and beliefs are understood as 'constitutive' of actions. In terms of causality, I explain why individuals respond to their world like they do[24]. Secondly, a different understanding of 'generalisability' is in place. Typically, generalisability describes the extent to which research findings can be applied to settings other than that in which they were originally discovered. It is imperative to distinguish statistical inference (mainly used in quantitative research) from case inference (Flyvbjerg, 2001; Yin, 2012). While statistical findings are mainly generalised to populations, interpretive work builds theoretical premises which function as tool to make assertions about situations akin to the one studied, with the help of in-depth analytical investigation (Yin 2012). This type of generalization is often branded as "analytic generalization" (Lincoln & Guba, 2000, p. 171; Yin, 2012, p. 18). In doing so, interpretive research unravels in-depth mechanisms of 'how' and 'why', that are more difficult to ascertain in large N-studies. In my interpretive research this is achieved by providing sufficient thick description of these mechanisms so that others can assess how plausible it is to transfer insights

[23] This is sometimes called "Sherlock Holmes causality", due to the careful mapping of clues in a specific context and the tracing of connections among events (Yanow & Schwartz-Shea, 2006, p. 108).

[24] Rather than explaining why an event A would immediately lead to an event B (sometimes called 'billiard ball causality').

from that research study to another setting. It enables others to build on research insights they find trustworthy (Lincoln and Guba 1985). For this reason, in the following chapters, I always sufficiently contextualize my data, and connect it to specific actors and settings. As such, interpretations are embedded in, rather than abstracted from the settings of the actors studied[25].

HOW? GENERATING AND ANALYSING DATA

In what preceded, I noted that my evidence is generated, rather than collected (or 'mined'). In what follows I will say a little more about how I generated this data. First, I will substantiate my case selection and choice of respondents. I will also discuss my access to the field, and reflect on my position as a researcher. Secondly, I will discuss my intertwined process of data collection and analysis; elaborating especially on the abductive logic followed.

CASE SELECTION: WHY THE STRESS-TEST?

My case of a performance indicator is the EU-wide stress test(s) of the banking system, conducted by the European Banking Authority (EBA) in close collaboration with the European Central Bank (ECB). The stress test projects how big, systemically important, European banks would perform in a stressful

[25] Some extreme relativistic views, espoused under the banner of social constructionism, have indeed led down a road to paralysis. 'Hard core' postmodernists sometimes claim data cannot (and should not) mean anything beyond the context of an interview. They tightly cling to the claim that "there is no truth" or "we cannot say anything". Although I do not support any claims to generalisability in a positivist sense, I do believe my findings can inform beyond the research context in their own way.

macro-economic scenario. The result of the stress test is a listing of the banks, displaying their initial capital ratios and their projected capital ratios after the stress scenario. These capital ratios can readily be used to rank banks according to their performance (high to low percentages). In early rounds of the stress test (up until 2011), a benchmark was set (at an 8% ratio), discerning which banks "passed" or "failed" the stress test (EBA, 2019a).

I have opted for the EU-wide stress test(s) for three main reasons. I briefly discussed these in the introduction, and I will reiterate them here. First, and most importantly, finance as a policy domain combines strong technocratic institutions and established professions (contributing to depoliticization of the sector), with high political stakes and strong political agendas (hinting to repoliticization) (Jessop, 2014). This makes it a good case to consider both political and technocratic drivers in knowledge production. Secondly, it is a recent indicator. The first round was conducted in 2009 and it has changed every subsequent year. As the objective is to study how performance indicators are made, and the effects of this design, it is useful to study a 'novel' indicator still in full development. A lot of disputes regarding design choices still remain unsettled. As multiple iterations of indicator development are finished, it is interesting to follow the modifications over the years and the different mechanisms at play. Moreover, this makes it more practical to speak with respondents involved in the indicator's design, compared to more longstanding indicators such as GDP. Finally, little socio-political research has been done regarding performance indicators in the financial sphere (an important exception is Farlow (2015)), despite their societal salience, especially in the aftermath of the 2008 crisis and accompanying public bail-outs.

The study is an embedded case study. Six EU-wide stress tests (2009, 2010, 2011, 2014, 2016, 2018) have been conducted so far, each iteration introducing alterations to the design of the performance indicator. Conducting a case study will enable a fine-grained understanding of the different dynamics at play in the making of the EU-wide stress test. The essence of a case study is that it offers a means of investigating complex social units, anchored in real-life situations. The data I will be able to collect in a case study will be a lot richer and of greater depth than would be obtained by other research designs (Yin, 2012).

SELECTING RESPONDENTS: MAPPING FOR EXPOSURE

I am in search of meaning, rather than law. Because I am not looking for general laws that hold for an entire population, it would not make sense to draw probabilistic samples or make a purposive case selection. What I did instead was map for exposure[26]. This means identifying different kinds of people, in different positions, with different roles and different understandings. I then aimed to maximize this variety, to expose myself to a wide array of meaning structures. I planned to speak with actors on both sides of the performance indicator, both regulators (ECB, EBA, National Bank of Belgium) who designed the stress test and regulatees (banks) whose performance was measured. As consulting firms also play an important role, aiding both sides in conducting the stress test, I included them as well.

[26] I have chosen to give up the rhetoric of the term 'sampling' as it originates in the probability requirements of inferential statistical science; it is a technical term that refers to the scientific possibility of generalizing from a sample of a population to the population as a whole, within some degree of certainty. This is not possible nor desirable in my research design.

For practical reasons of access, I chose to conduct interviews in Belgian banks. Of course, the national context can have a large impact on how people experience and make sense of the stress test exercise. This is not problematic, but it should be considered that these processes of meaning making might differ in other countries. However, this national angle was broadened through interviews with consulting firms who aided banks in multiple countries. In these interviews I was able to assess to what extent Belgian experiences and processes of meaning making were recognised and shared across European banks. Interesting in this respect is that consulting firms often had large-scale studies and reports (based on surveys) summarizing the general sentiment of the EU banking community towards the stress-testing exercise. In my in-depth research I was able to unpack these sentiments during interviews and understand where they came from. Moreover, my interviews with both Belgian and European supervisors allowed for a broad perspective.

To select respondents, I first contacted the chief risk officer (CRO) in each Belgian bank involved. I snowballed from there on, asking the CRO to identify other key actors in the bank. I also asked respondents which consulting firms they worked with, and if they could put me in touch with their contacts. I then looked up the departments at the ECB, EBA and National Bank of Belgium (NBB) in charge of the stress test. I additionally looked at ECB publications on the EU-wide stress test, and contacted the authors. I planned a research visit in Frankfurt and London, in order to be able to conduct face-to-face interviews at the ECB and EBA offices.

Selecting respondents is one thing, gaining access to them is another; especially when dealing with elites (Lilleker, 2003). As Abolafia (1998, pp. 78–

79) writes about financial elites: "Like other elites, they are insulated from observation and protective of their time. The researcher must often pass through several levels of gate-keepers to gain access and may be rebuffed at any level." I found it a fruitful strategy to begin with finding contacts in Belgian banks. As high-level respondents have limited time to spare, I found that it was useful to find an angle that showed them how my research, and their time, was not only of use to me and the scientific community, but also served their interests. In this case, desk-research showed that banks were very frustrated about the stress-test, but felt like their concerns were not being heard. I offered a listening ear and a means to put their concerns to paper in my dissertation – anonymously of course. I also promised to present my research findings to them at the end of the run, and discuss their relevance for the bank. I found that with enough phone calls, LinkedIn messages, and reminder emails, in the end, everyone I set out to interview in Belgian banks, accepted my request.

Gaining access to the ECB was a more difficult task. ECB employees are officially not allowed to give interviews; all communication is supposed to go through the press desk. As such, I took a more bottom up approach here. I looked for ECB publications on the stress test that seemed relevant to my research, and contacted the authors; asking them to discuss their article and speak a bit more about the stress test in general. The researchers I contacted all accepted my request. I asked them who the other relevant people to speak with were, and asked them to put me in touch. I was able to speak to two high level employees; whom, for purposes of anonymity, I cannot say much about. Only one respondent at the ECB abruptly decided to end our

conversation after about five minutes, as he did not feel at ease breaking the rules and speaking with me. I did not press the respondent on this, but simply thanked him anyways and wrapped up. Gaining access to the EBA was not as problematic as I expected. Perhaps because I mentioned that I had already interviewed respondents in banks as well as the ECB, and as such they perceived my research as already more legitimate and worth their time. Contacting respondents at the NBB was similarly unproblematic, after a few emails back and forth explaining my research in a little more detail, the people I contacted accepted my interview request.

I continued interviewing until I reached a point of saturation[27]. I conducted fourty-five conversational interviews with thirty-three people. The interviews on average lasted seventy-five minutes. We did a first round of interviewing in banks and consulting firms in 2015/2016, a second round of interviews was done at the ECB in 2017, and a in third round in 2018 I revisited banks and consultants, and additionally spoke to respondents at EBA and NBB. A full list of anonymised interviewees and dates of interviews can be found in annex. All interviews were done face-to-face and recorded digitally, except for two interviews at the ECB where only note taking was allowed.

REFLEXIVITY & POSITIONALITY

Where positivist research strives to reduce the influence of the researcher to an absolute minimum, striving for replicability; this is not a direct concern for interpretive work. Rather, it is acknowledged and embraced that the

[27] I frequently revisited my interviews to check whether they still touched on new themes or brought forth new information.

researcher, as the instrument of coding and analysis, has an important impact on data generation (Schwartz-Shea & Yanow, 2012). As such, self-reflection and reflexivity are important elements of the research process. For instance, as all my respondents were (often elder, male) experts, I was aware that they might be less likely to take a young, female, social scientist seriously. As such, I read up extensively on financial risk management, to acquaint myself with key terms and abbreviations. I also dressed business formal, as to fit into the working environment. I do not have an economic or financial background, so – despite my extensive desk-research, a lot of 'givens' for my respondents were not at all obvious to me. I always made my background in social sciences clear, albeit noting my interest in the financial sector. In the end, I think this was actually an advantage, because respondents took a lot of time to explain different processes and operations in their organization, providing me with thick contextual descriptions to analyse. It was also an advantage that given my lacking background in finance, I had no preconceived opinions of the stress test or how it was designed, whether it was good or bad. This left me open to hear many different interpretations and viewpoints, rather than (subconsciously) looking for evidence that confirmed a previously held attitude.

GETTING DIRTY WITH DATA, ABDUCTIVELY

Research practice often begins with the identification and definition of concepts, followed by operationalization in the form of variables, to result in hypotheses that establish relations among them. Underlying this is a

particular orientation towards knowledge and its sources, that does not harmonize with interpretive research.

Rather than a deductive rationale, that begins with a rule and looks at cases to either confirm or falsify it, interpretive researchers often prefer an inductive approach that starts from the data. However, the commitment to a fully inductive approach creates an epistemological and practical dilemma (Timmermans & Tavory, 2012). Researchers are expected to generate new theory without being prejudiced towards any existing (pet) theory, but still they are required to be broadly familiar with a range of existing theories to make sufficient abstractions from their data to generate new theory. A solution to this conundrum is provided by Peirce's notion of abduction (Fann, 1970), further refined by Timmermans & Tavory (2012). The etymology of abduction suggests a 'leading away' from old insights by puzzling research evidence. Abduction has a logic distinct from deduction or induction. Abduction starts with puzzles and seeks to explicate them by identifying conditions that would make them less perplexing. It suggests an iterative back-and-forth movement between data and theory. I take the stance that there is no 'view from nowhere'. Without prior knowledge, without some prior 'conceptual boxes' (Kuhn, 1962), I could not organise all of the stimuli that came at my senses; I would, in a cognitive sense, be blind to them. To foster this connection between theory and data I choose sensitising concepts[28] over formal hypotheses. Interpretive research rarely proceeds from hypotheses,

[28] I use sensitising concepts, rather than definitive ones, as they suit my methodology better. Where definitive concepts provide prescriptions of what to look for, sensitizing concepts suggest more general directions along which to look. This allows more space to generate data (Bowen, 2006).

because the researcher does not know ahead of time what meanings will be found.

Practically, this means that I began my data generation by conducting a document analysis of the official publications on stress-testing found on the website of the European Banking Authority (EBA 2014). The EBA publishes official documents regarding why the stress test is conducted, the methodological guidelines for conducting it, and the official results of the exercise. I used these official documents to gain a baseline understanding of how banks' health is constructed publicly. I also looked up scientific publications on stress testing and read several books[29], selecting the most relevant paragraphs and collecting them in separate Word-files. I uploaded these documents to NVivo and coded them. I then tried to situate them within the literature on Public Administration (PA) and Science and Technology Studies (STS), as briefly summarized in the previous chapter. I used these insights to draw up a semi-structured topic guide to take with me during my first round of narrative conversational interviews in Belgian banks and consulting firms. The aim of these interviews was to posit my expectations based on my desk-research and expectations from PA and STS literature vis-à-vis how the respondents in the field experienced the process of performance measurement and its implications. I coded these interviews and sought new theoretical insights to address puzzling findings.

[29] I used approximately thirty scientific articles on stress-testing, and book-wise drew mainly on Mario Quagliariello's (2009) 'stress-testing the banking system' and Timothy Geithner's (2015) 'Stress test: Reflections on Financial Crises'.

I then conducted a second round of interviewing, to add the supervisory point of view (as a negative case[30]) to the overall understanding of the stress testing exercise and bring in alternative insights. Again, I coded these interviews and recalibrated my sense making of the field. Finally, I returned to a number of the respondents to further address and challenge theoretical expectations; this member checking served as respondent feedback and respondent validation for my findings (Schwartz-Shea & Yanow, 2012). At the end of this process I was able to address my research questions with novel theoretical insights; speaking to both PA and STS literature.

Though intertwined, for purposes of clarity, I discuss the data generating and data analysis stages separately in more depth below.

INTERVIEWS

Among the multitude of methods to generate data[31], is the ability to talk with people: the interview. Interpretive interviewing is intended to explore the meanings of events. They are often described as 'purposive conversation' (Yanow & Schwartz-Shea, 2006)[32]. Conversational interviews[33] can enable the exploration of how people make sense of their experiences and how this sense making connects to action. It is important to note that interpretive

[30] Including negative cases means searching for and discussing contradicting patterns or explanations than those so far emerging from data analysis (Schwartz-Shea & Yanow, 2012). It proved very interesting to be able to see both the point of view of the supervisors, and the regulated entities.

[31] Constructivist ontology and interpretive epistemology lead me to see knowledge being generated by both participants during the interview.

[32] Purposive because the open-ended interviews do not ramble all over the place, without structure or direction. The interviewer directs the conversation in a certain way.

[33] The original latin meaning of conversation is 'wandering together with' (Kvale & Brinkmann, 1996), which ties in with the 'traveller' (vs. miner) metaphor.

researchers are not 'trapped' by what people tell them, or by prejudice. They are alert to the possibility of partial knowledge and multiple perspectives. These are not avoided or 'controlled for', but they are acknowledged, engaged and analysed.

To conduct my interviews, I used a semi-structured[34] topic guide rather than a fixed set of questions. This allowed me to discuss the same set of topics with all respondents and still act in a responsive way. A specific script could end up imposing my own framing, whereas a topic guide leaves more room to incorporate terms, concepts, language and behaviour used by participants (Yin, 2012). The topic guide used for interviews in banks is included in appendix 1, topic guides for other respondents were very similar. My topic guides were built up according to the abductive logic I described earlier. My topics are based on my sensitizing concepts, which were derived from an iterative back-and-forth movement between literature and data[35].

The topic guide was just that – a guide. This left room for respondents to play a part in the ultimate course of the interview. Sometimes, after introductions, the respondent dominated the conversation by discussing what he or she thought would be of interest for my research (and usually these matters were very interesting). On those occasions I disregarded the topic list and let them speak freely first. If I felt certain important topics were left unaddressed, I made sure to pick up on those towards the end of the

[34] 'Structured' meaning that I planned out different stages of the interview. This is not to say that the entire order is planned out. It leaves room to move between topics in the same stage.

[35] As such, the topic guide has been altered along the way, especially after my first set of interviews. Revisions included a change in the order of topics, and eliminating and adding new topics. The first interviews served as a pilot of the topic guide. This data was not excluded from my data set since interpretive research, unlike positivist research, does not require standardized data.

interview (my rule of thumb was to steer back to the topic list around the last 1/3d of the conversation). Moreover, some interviews were shorter than others, due to time pressure. I didn't discuss all topics with all respondents[36]. Different people were more knowledgeable about different things and involved in different parts of the process. I made a list of what I needed to know and who would be most likely to know this, but I also gave my respondents space to discuss what they deemed most relevant.

Another point worth mentioning is the process of transcribing interviews and taking notes. I chose to do my own transcribing for confidentiality purposes and to secure the many details relevant to my analysis. I transcribed the interviews verbatim, retaining repetitions, sighs, pauses, emphases in intonation and other meta-data as much as possible[37]. This to aid later interpretation of the data. I tried to keep my note taking during the interview to a minimum, this to stay focused on the actual conversation and to avoid distracting my respondent. I made a habit of adding to my notes immediately after the interview or while doing transcriptions. I wrote down all and any thoughts that occurred to me in notebooks, on post-its scattered across my desk, in numerous word-files, and in my phone during the train ride home. I regularly brought all these notes together physically in a memo file. I jotted down passages of interviews that reminded me of a certain book or article, possible interview questions,

[36] Although I did attempt to. I often arranged follow-up interviews and made sure to pick up on things that were left out previously.

[37] I consciously use the phrasing 'as much as possible'. Transcription is always an interpretive process, where the differences between oral speech and written texts give rise to a series of practical and principal issues. Transcriptions are translations from an oral to a written language. So, what is said about translaters, also applies to transcribers: traduire traittori – translators are traitors.

important words, comments on the content of the interview, or how I felt the interview went. I also made notes of good quotes that caught my ear during interviews.

During interviews respondents sporadically referred to (and even provided me with) additional textual material. This included methodological guidelines, power point presentations, internal and external e-mail communication, reports and publications. Although this data was limited in size, it provided additional insight into the respondents' narratives. I thus coded this material along with my interview data.

CODING & ANALYSING

I uploaded these transcriptions along with my research notes to Nvivo to manage and analyse my data systematically. I began with emergent coding close to the text (Drisko & Maschi, 2015), in order to stay true to respondents' representations of events, and the context in which the data was generated. I then compared codes across interviews and grouped them together in overarching codes to establish sensitizing concepts (Bowen, 2006). I related these concepts to theoretical notions that I used to inform subsequent rounds of interviewing. My code book, that links my emergent codes, to sensitizing concepts and theoretical notions, can be found in appendix 3.

My analysis started very early on. The first step was to look at my pilot interviews and make sure my project made sense to my respondents. Once they confirmed that the stress test was indeed a highly relevant and highly problematic indicator, I could go on to the next step: recognition. In this

second step I looked for topical concepts, themes and events in my interviews. I read, reread and reread again, moving back and forth abductively between data and theories until the interview data started to make sense in a new way. I kept a clear audit trail, taking note of all the steps I took from the start of my research to the end; who I interviewed and why, thoughts on why I used certain codes, and how I linked different concepts together. Data analysis is inherently a shaping of reality, rather than an exact point-for-point recapitulation of data. The analytical process entailed classifying comparing, weighing, and combining material from the interviews to extract meaning, reveal patterns, and stitch together descriptions of events into a coherent narrative (Coffey & Atkinson, 1996).

In Conclusion: what my data can and cannot do

Where many a reader (and writer) might regard the methodological chapter as a necessary interruption to legitimate a far more interesting story, I see this differently. Rather than just a legitimation, a methodology provides a deeper insight in the nature of the data obtained and analyses. It helps us understand what our data can and cannot do or say. The philosophical positioning I began this chapter with serves as a conceptual frame of reference to clarify the nature, the strengths, and the weaknesses of the data obtained during my interviews. I would like to highlight two of the key features of this data.

First of all, knowledge is produced. This means that knowledge is not 'mined', but socially constructed by interviewer and interviewee. The events I describe in the following chapters are not 'the truth about those events', but

an interpretation of an interpretation of those events. Stories are always told, retold and interpreted from somewhere. As such, interpretive conflicts are quite common. I did member checks where possible to make sure that I did not completely miss the point, and included negative cases to ensure a variety of viewpoints and interpretations. So, this is not 'what happened', or 'what they said happened'. This is my take on what happened, informed by both a specific body of academic literature and empirical evidence.

Secondly, knowledge is contextual yet generalisable. I was not interested in universal truths that I could automatically abstract from the studied setting. I was looking for processes of meaning making that were embedded in the context. Because of the heterogeneity of contexts, translation between contexts becomes an issue. The findings in the following chapters are not automatically transferable to other situations (statistical inference). This is not to say that nothing can be learned from these contextualised findings. Much can be learned over settings and contexts beyond what is similar. In this interpretive work I build theoretical premises which function as tools to make assertions about situations akin to the one studied, with the help of in-depth analytical investigation (case inference). Instead of generalisable laws, this research focusses on unravelling patterns of causality. Carefully mapping and connecting events, to further an understanding of how certain events were triggered in one context, and subsequently, how they might be triggered in others.

How indicators measure:

The incommensurables

This chapter is published in a slightly adapted form as:

Kempeneer, S. & Van Dooren, W. (2019). The incommensurables: the arduous art of making a regulatory indicator. *Critical Policy Studies*

Chapter overview

In this chapter I follow Michel Callon and Fabian Muniesa's (2005) idea that in order to be measured, entities must be made measurable, and comparable first. This is done through the process of commensuration. Despite a substantive literature, little empirical work has been done to further our understanding of the social and political processes that drive commensuration. I use Actor-Network Theory (ANT) to enrich the existing literature with an in-depth account of how commensuration is negotiated. I find that despite a preference for commensuration, regulators allow 'incommensurable' categories to exist due to largely unrecognised regulatory benefits, such as learning opportunities and innovation. This also shows that measurement 'bias' can be intentional, and is not necessarily the product of capture, gaming, or industry lobbying.

HOW BANKS ARE MEASURED

On a crisp October day in 2013, 130 banks received a letter from the European Central Bank (ECB). Congratulations were in order; the banks were selected as 'significant financial institutions' in Europe. As such, they would fall under the ECB's Single Supervisory Mechanism as of November 2014. This meant that from now on they would be reporting to the ECB, instead of their own country. But before being admitted to the top league of European banking, the banks had to prove that they were healthy. To do so, the ECB would conduct a 'Comprehensive Assessment', scrutinizing all 130 banks to test their financial resilience. A key part of the Comprehensive Assessment was a stress test. The ECB announced it would be the toughest stress test to ever be conducted in the European financial sector. A top official allegedly said[38]: 'There will be blood'.

The stress test assesses how well banks would cope during a three-year crisis scenario. More specifically, the test shows how much capital a bank still holds against its risk weighted assets after three years of crisis (projected adverse capital ratio). Table 3 below shows an excerpt of the results (EBA, 2014). Even without a good understanding of how the indicator is made, it is possible to infer which banks are healthy (high percentages) and which are not (low percentages). The results made headlines globally. In the run-up to the stress testing exercise banks were seen to raise significant amounts of capital to increase their score, conscious that financial markets would be judging their results (Titcomb, 2014). In subsequent rounds of stress testing,

[38] According to respondents, Daniele Nouy, the chair of the Supervisory Board at the ECB, made this statement while announcing the stress test.

the indicators were also formally linked to regulatory interventions, such as banks' Supervisory Review and Evaluation Process (SREP), which amongst others determines how much capital banks are legally required to hold.

TABLE 3: BANK SPECIFIC RESULTS OF THE STRESS TEST

Country	Bank	2013 capital ratio	baseline 2016	adverse 2016
AU	BAWAG PSK	14.3%	11.9%	8.5%
AU	Erste Group Bank AG	10.0%	11.2%	7.6%
BE	AXA Bank Europe SA	14.7%	12.7%	3.4%
BE	Belfius Banque SA	13.5%	11.0%	7.3%

Regulatory indicators, such as the stress test, pervade transnational governance. Indicators and rankings play an important role in the way governmental and non-governmental organizations distribute attention, make decisions, and allocate scarce resources (Rottenburg & Merry, 2015). The OECD's Programme for International Student Assessment (PISA) steers decision making in education policy (Gorur, 2016); the Human Development Index informs the United Nations Development Program (Davis et al., 2012b); and the EU's Open Method of Coordination helps policymakers monitor and measure progress in various policy domains across member states (Marlier & Atkinson, 2010). We define an indicator, following Davis et al. (2012: 75), as:

> "A named collection of rank-ordered data that purports to represent the past or projected performance of different units.

> The data are generated through a process that simplifies raw data about a complex social phenomenon. The data, in this simplified and processed form, are capable of being used to compare particular units of analysis (such as countries or institutions or corporations), synchronically or over time, and to evaluate their performance by reference to one or more standards."

The regulatory character of such indicators should be emphasised, as they can steer behaviour even without legal enforcement. The reputational pressure that indicators exert with their potential to rank performances, is often effective in in securing compliance or behavioural change (Grabosky & Braithwaite, 1986; Tervonen-Gonçalves, 2012; van Ostaijen & Scholten, 2017).

WHY CARE ABOUT COMMENSURATION?

Several reasons present themselves to study this process of how transnational indicators are made. First of all, commensuration, the need to homogenize across contexts, reaches its pinnacle in the European context (Bruno, Jacquot, & Mandin, 2006; Mügge, 2016). Despite the diversity in national contexts, Member States, and their policy issues, are intricately connected in many ways, calling for an overarching European regulatory system (and legislation) (Jurgen Habermas & Derrida, 2003; Kohler-Koch, 1996). This became especially clear in the financial crisis where the interconnectedness of the banking system caused risk to spill over national borders and contaminate the entire European sphere (De Bruyckere et al., 2013). The complexity of transnational policy issues simultaneously enhances both the

need for clear-cut indicators to regulate, as well as the difficulty to create such indicators across varying contexts. As such, efforts towards, as well as struggles with commensuration will likely play a pivotal role. Discussions on transnational commensuration and standardization tie in with Barry's (2012) work on transnational knowledge controversies. One of the critical difficulties in governing transnational issues, such as financial policy, is a lack of transnational consensus on matters of fact, or how evidence should be interpreted.

Secondly, Welsh (2017) calls for a more critical and political analysis of rankings. Likewise, Mügge (2016) explicitly calls for political scientists in particular to pay more attention to the forces that determine indicators' design. Since the 1960s, political scientists have used theories of regulatory 'capture' to explain regulatory outcomes and designs. It conveys a sense of illegitimate expropriation, performed by one powerful group over others (Baxter, 2011). A substantial body of literature studies the privileged interactions between industry and public authorities (Bunea, 2013; Hanegraaff et al., 2016; Klüver, 2013; Lowery, 2013). When indicators paint a predominantly positive picture of regulated entities, this is often ascribed to regulators designing a less critical performance measure (Woll, 2014). In the case of the ECB's banking stress test, such accusations were echoed in public debate as well. The crisis scenario that the ECB banks were subjected to was deemed far less severe than that of the Federal Reserve or the Bank of England, leading to misleadingly positive results for the European Banking Sector (Cecchetti & Schoenholtz, 2016). Beyond 'capture', public administration scholars criticize indicators for their susceptibility to 'gaming' (Bevan & Hood, 2006). Aware of which behaviour would be measured by the

indicator, organisations or countries are seen to manipulate performance outcomes to their advantage, painting an overly optimistic picture. However, it is yet to be studied how these mechanisms come into play in designing regulatory indicators.

In what follows I build on literature in quantification (Desrosières, 1993; W. N. Espeland & Sauder, 2007; W. N. Espeland & Stevens, 1998; Hacking, 1990; Peeters, Verschraegen, & Debels, 2014; Porter, 1995) to improve the understanding of how indicators are designed. This is increasingly important, because they affect our understanding of the world in subtle and often unrecognised ways (Rottenburg & Merry, 2015). As such, before we can use these numbers in policy and public debate, we need to understand precisely where they come from. I draw on my interview material with with stakeholders in risk departments in Belgian banks, consultants, the ECB, the EBA and the NBB to better understand how the stress test measures, and makes banks measurable.

I find that although these design choices are politically motivated, there is more to the story than mere gaming or regulatory capture. Introducing incommensurability can be seen as a manifestation of a critical epistemological attitude that is objected to a rationally calculable reality. In the concluding section, I reflect on what these findings mean for broader academic and public debate, relating it to the Frankfurt School's critique of instrumental reason (McCarthy, 1990) and more specifically to Habermas' notion of communicative rationality as an alternative to instrumental rationality (Habermas, 1990).

MEASURABILITY AND COMMENSURATION

Commensurating systems, means taking diverse qualitative systems and homogenizing them on a common metric, facilitating comparison (Espeland & Stevens, 1998). Literature on commensuration pays particular attention to the socio-political forces that lie behind this standardization process. Regulatory indicators used in transnational governance, commensurate systems over national contexts. They assume that is it possible and desirable to compare complex systems across countries according to uniform measures. In our case, the ECB stress test aims to make a standardized comparison of banks' risk across Europe. Commensuration literature stresses that things 'are' not comparable, but that they need to be made comparable. It thus proves interesting to pay special attention to how this commensuration process feeds into transnational indicator design, and how indicators make diverse systems comparable.

Commensuration is a social process. It begins with the idea that it is meaningful to compare a set of things. For example when estimating the costs of a large infrastructure project, we now believe that it is important to take into account the potential loss of natural resources, and other costs to the environment (Vickerman, 2007). To do so, we need to find a way to compare these costs and benefits. This is often done by putting a price on the loss of land, or the quality of air, that we then can compare to the cost of traffic congestion or employment. However, practically finding ways to value and compare diverse inputs is no small feat (Patterson, 1998). As such, commensurability is not only a social construction, but a social accomplishment that requires substantial efforts. Entire agencies, industries

and even disciplines are dedicated to finding ways to compare (the value or performance of) different systems (Jasanoff, 1986). In our case, it was only after the crisis that the urge arose to create a standardized pan-European comparison of banks, before this, each National Competent Authority had its own method of assessing banks' health. These different national strategies were then also tailored to the local context and banks' business models. Finding a way to compare these banks on a Level Playing Field across Europe was thus a challenging task for the stress test. Especially when it comes to big, systemically important, banks. Standardised reporting requirements can make it seem as if banks are easily compared across contexts. However, when you take a closer look, the financial products that these big banks hold are so disparate and complex, making it very difficult to calculate and quantify their value and risk. How much an asset in a big bank is worth, or what its risk is, can be an as vexing question as how much a human life is worth (at least in terms of all the different parameters that can be considered[39]). A wide range of qualitative properties of the asset need to be transformed into quantities, and this can happen in a variety of ways, according to various assumptions, theories and models.

THE INCOMMENSURABLES

Just as things can be seen as commensurable, they can also be seen as 'incommensurable', or undesirable to compare. For example, Ackerman and Heinzerling (2005) argue that it is morally wrong to compare the value of

[39] And when you hear experts talking about this it often seems as if there is as much at stake.

human lives, health, or the environment to economic gains, especially not by reducing them to 'cold dollars'. These things are seen to be 'priceless', and thus incommensurable. Incommensurability is as much a social construction as commensurability, and also requires work (W. N. Espeland & Stevens, 1998). Incommensurability claims are often supported by moral arguments, that for example, human life should be valued above anything else, or the environment should be protected at any cost (Ackerman & Heinzerling, 2004). In the case of the stress test, banks argue that it is not fair to compare one portfolio to another at face-value. They claim that their assets are simply too different in nature to be assessed according to the same standards.

A more substantive critique comes from the Frankfurt School, who condemn the wider economic, political and social effects of commensuration (for a good overview see Smulewicz-Zucker, 2017). In the Dialectic of Enlightenment for instance, Horkheimer and Adorno (1947) discuss the standardising effects of mass media, akin to a 'culture industry' producing uncritical identical individuals, locking in power relations, and reproducing dominant discourses. In a similar fashion, indicators (and other technologies) can act as a standardising instrument for control and domination (Marcuse, 1941). Treating banks as incommensurable would then be the only way to allow critical rational debate to triumph over manufactured information. Here incommensurability is not defended from a moral point of view, but from an emancipatory one.

Regulators are typically in favour of commensuration (Gorur, 2016; Porter, 1995; Scott, 1998). The illegibility of local contexts is an administrative headache for regulators. Without comparable units of measurement, it proves almost impossible to monitor, compare, or regulate performance in

various policy domains. Regulators need tools like standardised indicators to understand and manage the large and complex reality. On the other hand, regulated sectors typically fight these commensuration efforts. They see their unique qualities stripped away and do not feel accurately represented by these standardised measures. They often have their own distinct view and contextual interpretation of the categories that are taken into account (Peeters et al., 2014)

STUDYING COMMENSURATION WITH ANT

In what follows I analyse this process of commensuration empirically, with the case of the ECB's banking stress test. To do so I use a framework borrowed from Actor-Network Theory (ANT), Callon's (1984) 'Sociology of Translations'. A basic assumption of ANT is that everything we see in the world is built up from a set of relationships between people and things, a so called 'Actor-Network' (Latour, 1999; Law, 1999). Callon uses the notion of 'translation' to explain how an Actor-Network, comes to be represented by a single thing, in our case an indicator of health. The linguistic metaphor of translation emphasizes the manner in which interests, goals, or desires are represented, simplified, and transformed in the production and mobilization of artifacts.

The logic of translations is helpful in studying processes of commensuration. Callon (1984) distinguishes four key moments of translation: problematisation, interessement, enrolment, and mobilisation. To untangle the concept of (in)commensurability, it is helpful to unpack it according to the various stages. For instance, during problematisation we can

look into why and for whom commensuration is necessary, interessement and enrolment give us an understanding of how commensuration is negotiated in practice, and mobilisation can help us understand how commensuration ties into wider societal processes. This analytical framework thus gives us a more multi-facetted and critical understanding of commensuration and claims of incommensurability.

COMMENSURATION IN ACTION

I will structure this empirical section using Callon's four key moments of 'translation': problematisation, interessement, enrolment, and mobilisation (Callon, 1984). Looking at commensuration as a process of translations, gives us a more in-depth understanding of the role of incommensurables. The inability to commensurate is often written off as a failure in a process, rather than a deliberate action. Throughout the different translations, we gain a better understanding of what drives (in)commensurability and how.

PROBLEMATISATION: SHOULD BANKS BE COMMENSURATED?

In the first translation, problematisation, a given actor analyses a situation and provides a specific problem definition. An important requirement is that this problem definition, and the subsequent proposed solution, rings true to other actors: What is the problem that needs to be solved?

In the case of the stress test, the 2008 financial crisis made clear that something was wrong in financial regulation. But what? National supervisors,

along with the rest of the world, failed to see the crisis coming. A risk director in a bank noted:

> "Before the crisis a lot of banks used quantitative models (…) but then people saw that a lot of banks failed and had issues during the crisis even if the models said everything was ok."

This illustrates how old indicators of banks' health showed significant shortcomings during the crisis. Moreover, there was a rising distrust towards national regulators. They were said to be 'captured' by the banks, no longer safeguarding the public interest but preoccupied with national interests and 'their banks' looking good. Banking supervision was concentrated at the national level before the crisis, because national authorities stood closer to banks and were seen as better equipped to understand the complex legal and socio-political context they operated in. A risk expert noted:

> "You cannot take an asset in this bank, and just compare it to an asset that might look the same in another bank. You have to understand what's underneath this asset. For example, credit quality, we have clients with good savings accounts. That's typically Belgian actually. This is going to affect the PD [Probability of Default] and the LGD [Loss Given Default] and such."

This shows that comparing banks is not so straightforward. Before the crisis banks across Europe were treated as 'incommensurable' from a regulatory perspective: Banks were seen as too different in terms of business models, activities and portfolios, to be compared in a uniform way. Although there

were general guidelines, banking supervision differed across countries. This changed after the crisis. The crisis revealed how interconnected banks in Europe were, and that despite their differences, it would be meaningful to compare the systematically important institutions. The problem shifted from a localised national issue, to an integrated European issue. This led to the establishment of the Single Supervisory Mechanism (SSM) in 2014, that transferred supervisory power to the ECB. And as such, finding a way to commensurate these different, yet interconnected European banks became a key priority for the stress test.

Interessement and enrolment: To commensurate, or not to commensurate?

The next translations are interessement and enrolment. In interessement actors seek to lock other actors into specific roles. Actors, in ANT can be both people and things. To make this interessement successful, Callon uses the concept enrolment: these are the negotiations that accompany the interessements and enable them to succeed (Callon, 1984). Think of interessement as writing up a script, and enrolment as negotiating everyone's part. In this section, I discuss the different actants that were 'interested' and 'enrolled' into the stress test, and the role they played in commensurating banks. In the stress test the process of interessement and enrolment embody how the ECB wrote up the methodology (or script). By looking at which rules (or roles) were instated and how, we can understand how they play into processes of commensuration and claims of incommensurability.

Commensuration at any cost?

The starting point for the stress test is banks' balance sheet. Balance sheets give an overview of banks' assets and liabilities, these can be categorised in different sub-groups with specific characteristics. The stress test projects the impact of a three-year crisis scenario on banks' balance sheet. In reality, banks would make changes to the items on their balance sheet during a crisis, such as selling off bad assets. However, in the name of the LPF, banks' balance sheets were kept static over the three-year scenario. Some banks might make bigger changes than others, and it would be hard to compare the end results. The management decisions that a bank would make were not 'translated' into the stress-testing exercise, because they were impossible to predict and/or standardise. Keeping a static balance sheet was thus the only way to keep banks commensurable. This came at a cost. As we saw during the financial crisis, banks are intricately interrelated. The only way to map these interrelations, and understand their effects, is through a dynamic balance sheet. A respondent at the ECB explained:

> "Say a bank's balance sheet is hit, maybe they're going to be selling off corporations, so that's going to spill over to other countries, maybe, who knows. But this is what we would like to know. But without a dynamic balance sheet, we cannot know."

A static balance sheet makes it possible to make a fair comparison between banks, but it stands in the way of predicting how a crisis would affect the banking system in its entirety. Here we see clearly that commensuration, and comparability, limits the questions that indicators can answer. An indicator design that focusses on comparing how banks react to a crisis scenario, will

not be able to thoroughly answer questions about how these banks, and the banking system in Europe, would react to the scenario. The choice for a static balance sheet is a pragmatic one, that has little to do with steering results in any direction. Respondents in banks and at the ECB agreed that a dynamic balance sheet would be more adept at gauging a bank's true risk. However, respondents understood the choice to keep the balance sheet static to facilitate fair comparisons.

A next important design choice is the adverse scenario that the banks have to face. The stress test was designed with a common scenario, meaning all banks would be subjected to the same macro-economic turbulences. This was – again - motivated by the LPF. The ECB wanted to compare banks under the same scenario to eventually assess which banks performed better or worse. At first sight, a common scenario would facilitate the commensuration of banks. However, a respondent at the ECB criticized this choice:

> "What you want is a similar degree of pressure applied to all banks. But that does not happen with this single scenario. Banks are complex institutions, and each bank is sensitive to different things – it's like putting the same weight on different bridges."

So, on the one hand, it would seem fair to expose the sample of banks to the same scenario, but on the other it would also be fair that each bank would be exposed to the same amount of stress. This hints at the complexity of the questions that commensuration raises: Is it possible or even helpful to compare how different banks react to a crisis? The result of the commensuration might have less to do with the actual health of the bank, and

more with inadvertent sensitivities that a bank might have to a specific scenario, or even sheer luck.

Another important design feature is the 'common methodology'. The common methodology describes in detail how banks are supposed to calculate and report the impact of the scenario on their balance sheet. This document was key in the commensuration process. It forced banks to enter their balance sheet data according to standardized reporting standards and implement common definitions and assumptions to calculate asset values and risks. All banks filled in identical excel sheets with hundreds of data points, in identical columns and rows, that could easily be compared. Moreover, caps and floors were added to data points to further restrict large divergences in results. We asked where these caps and floors came from, and if they favoured any particular bank. A risk expert explained:

> "No, I don't think they favour anyone. They really are just put in place to keep the results conservative, to make sure the results of the banks are not too far apart. They don't make any economic sense either, in my opinion. They just 'assume' for all banks that for example some results cannot be positive, or that you have to calculate something according to [a set of rules]. And in our case, we have some exceptions to these rules, and we can explain this. For some assets we provide an extra insurance or so. But we're not allowed to take that into account. We have to calculate everything the same."

We pushed the respondents in banks on their ability to influence the design in their favour, the response was:

> "Well we talked with [an association of banks] but the thing is, it's hard to align interests, something that would be good for us is, is not necessarily good for another bank. So, it's not like you can press on these issues together. Even within Belgium. We did make the same arguments on some points, I called [a risk expert from another Belgian bank] to ask how are you going to interpret this rule, so we took a similar approach there. But these opportunities are limited."

This shows that 'capture' becomes more difficult when the interests of a group are not aligned, and the group sees its interests as 'incommensurable'.

The common methodology with the caps and floors seems to point in the direction of more commensuration efforts by the ECB. However, the fact that there is a common methodology, and banks are allowed to calculate the exercise themselves is quite remarkable. The ECB could have just conducted the stress-test themselves and left banks out of it completely. This is called a top-down stress test, where the central bank uses their own data or data delivered by banks, to calculate the impact of the scenario on banks themself. Using top-down models would take away any leeway banks would have to manipulate or game the exercise. Given the importance of the LPF, this would not have been a strange decision. Yet, the ECB opted for a bottom-up stress test, where banks were given some freedom to calculate the impact of the scenario themselves (albeit according to strict methodological guidelines). The question is: why?

Claiming incommensurability

We can be fairly sure that the reason is not technical infeasibility. The ECB has already developed top-down models to calculate the impact of the scenario on a bank. They just choose to solely use them for benchmarking purposes. A respondent at the ECB explained:

> The supervisors use the top-down results as a benchmark to judge the banks' results. And then there is a back and forth process with the banks. They can try and explain why their results are different from what we expect.

In this, the design allows some room for incommensurability. Banks are given the room to argue that their assets are incommensurable with seemingly similar assets in other banks. They can explain why their risk should be calculated differently than risk in other banks. During interviews three reasons came up for designing the exercise bottom-up. The first had to do with banks being able to give the best representation of their risk themselves. A respondent at the ECB noted:

> "Well I think banks should manage their own risk, I don't think supervisors need to do this. We really, and this is important, we want to foster the development of banks' own risk management capabilities. It is normal that banks should have a much better understanding of their risk. That is a big argument for a bottom-up stress-test. (…) From banks, it is not just a wilful act to spin figures, it's important to have a close discussion with institutions."

A risk expert in a bank added:

> "You can't just take an asset class and treat it the same in different countries. When you discuss this in international committees sometimes it looks as if each country is just defending their interest. But for example, mortgage lending, in some countries you can give back the keys, in others you still have to pay, you are still liable. So historically the defaults on the loans are very low. So, there is an argument to be made that these banks do have safe portfolios and you should not apply the same risk weight globally."

This shows that supervisors at the ECB and banks agree that banks' assets should be treated as 'incommensurables'. Here, an accurate understanding of the risk of a bank's asset, is valued higher than the commensurability of the asset. It is agreed on that it is not always desirable to compare the risk of even seemingly similar assets. To be sure, we've seen that accuracy has been sacrificed before in the name of commensuration. So, this still leaves us with the question of why the incommensurability is granted this time. Technically, it would be possible for regulators to run the whole exercise themselves. Yet, banks were allowed to use their own internal models. This can be better understood by the second reason that was given for granting incommensurability.

The second explanation for the bottom-up stress-test had to do with accountability and responsibility. Commensuration can shift responsibility away from regulated systems, onto regulators. A respondent at the ECB stated:

> "The team is great, but to do a top-down for all the banks in the comprehensive assessment, that's about 130 banks [in 2016]. I'm not sure that will happen. The inherent danger of using the same models for 130 institutions, that's a risk in itself."

A risk expert in a bank elaborated:

> "The regulators can do this top-down exercise. But I feel that supervisors are a bit apprehensive that if they do everything themselves, calculate it, publish it, then they are accountable. If a bank gets a good score but gets in trouble the year after, well the supervisor will be blamed fully."

This suggests that supervisors do not want to be completely responsible for assessing what will happen to banks' assets in a stress-scenario. Especially, because supervisors see a danger in banks using these top-down models, instead of developing their own models. A respondent at the ECB worded this carefully:

> "We never give all the information about our {top-down} models. Just enough to understand the model, but not enough to replicate it. We don't want banks to just take the models and use them. Then we would lose the bank-specific models, which are obviously valuable. We do not want a mono-risk culture. There is a big top-down, bottom-up discussion globally. It is good if banks use our models as inspiration, but to replicate, no. you want to keep some uncertainty, because you have model uncertainty. You cannot say 'this is the one model', you can't put all the eggs in

one basket. This should never become some unilateral guidance to banks."

This shows that even when commensuration is technically feasible, regulators choose to treat banks as 'incommensurables'. These reasons tie into Barry's (2012) notion of transnational knowledge controversies. Regulators cannot seem to choose one standard model to assess how banks' assets would perform under stress. There is no pan-European consensus regarding how much risk assets hold and how they would be affected under crisis situations. A consultant explained the trend over the past years:

> "The idea was at first to apply standard risk weights to standard asset categories. But as banks and their financial instruments became more complex, these standard risk models did not reflect the true risk anymore. A few years ago, there was a clear direction to more sophisticated models, giving more flexibility to the bank to develop internal models that would really make the bank able to simulate the exact risk of their bank and business model. But we've seen that banks are a bit using or playing with these models to go around the rules. So, we are seeing a clear trend towards more standardized models, and more simple models. Because the big problem with all these complex internal models is that you get results from the different banks and you cannot compare easily. When you see that for the same exposure, banks have different RWA [Risk Weighted Assets], that's not normal."

This demonstrates that regulators are struggling with building transnational knowledge, and a transnational consensus on how to gauge the risk of banks'

assets. In order to commensurate banks across Europe, transnational guidelines need to be established regarding how risk should be perceived and calculated. However, as each bank is developing its own complex financial products, it becomes even more challenging to create a homogenized understanding and calculation of risk.

In conclusion, this section shows us that on the one hand regulators go through a lot of effort to commensurate systems, but on the other that commensuration is no regulatory panacea. To be able to compare banks on a level playing field, concessions are made in terms of accuracy. However, commensuration is not pushed at any cost. Despite the possibility of designing a completely standardised top-down stress test, regulators consciously opt for a bottom-up exercise, treating banks assets as incommensurable. By allowing banks to use their own models, regulators create room for potential gaming efforts, where banks could try to make their portfolio look as good as possible. Designing a 'gameable' indicator may seem like a concession to industry interests, showing weakness from regulators side. However, this analysis shows that regulators consciously treat banks as incommensurable because of the regulatory benefits this entails, such as shared accountability, information access, and improved internal risk management capacity. Regulators are apprehensive in establishing transnational guidelines on how risk should be interpreted and calculated, because there are still controversies regarding the right way to calculate and measure financial risks. As such, if regulators were to establish a transnational norm, they would be fully responsible for the consequences in case the norm would prove erroneous. Moreover, they believe in the benefits that come

from leaving the debate open, and continuously reconsidering how risk should be understood and calculated. This can again be related to a more critical epistemological stance, where regulators choose to treat banks as incommensurables from an emancipatory point of view.

MOBILISATION: DESIGNING FOR RESULTS

The final step in Callon's framework is mobilisation: The actor-network starts to operate. This is the moment of truth, the final translation. Will the story hold, will all actors stick to their roles? The alliances made and consensuses agreed upon can be contested at any moment. Translation can become treason.

It is important to note that the result of the stress test needs to be able to 'hold' in the real world. It needs to be able to interact with other actors. As such, it's shaped by these possibilities for interaction. The stress-test is not created in a vacuum. In the mobilisation stage, it needs to be translated into the real world, and work there. A respondent at the ECB summarised this well with the sentence: 'you're stuck with what you can afford'. The scores on the stress test become a new fact and other actors will interact with this fact, they will make decisions based on this fact. So, you pre-emptively need to take into consideration what decisions these actors might make, and which decisions you can 'afford'. The respondent elaborated with a concrete example of how financial markets would react to the stress test, and which reactions the US could afford, and which reactions Europe could afford, he finished off stating:

> "So, it's obvious that the stress test results were mild. But this is not because the people who do these things are incompetent, or captured by banks, or are weak intellectually or whatever. There is also the dimension that it would have been irresponsible to come out with a cap request of 100 billion in such a situation [a conservative fiscal stance in Europe]. You're stuck with what you can afford."

A risk expert in a bank agreed:

> "It's politically motivated, an exercise like this. They know exactly how severe the stress will be on the banks. They developed it like that. Those hundreds of pages of rules, they know approximately what the outcome will be. They know perfectly, with the exercise they drafted now, they know that is the message they want to be spreading. And they do so in everyone's best interest."

As such, the message was clear from the beginning: the stress test would have to say banks overall were healthy. Indicators have important performative functions, they can become real in their consequences. If the ECB would come out with the message that banks were unhealthy, financial markets would have reacted this, only aggravating the situation. As such, the design of an indicator does pre-emptively need to take this performativity into consideration, especially when making results public. We confronted respondents with this. If the results were decided on from the beginning, why did we need such an elaborate exercise? Here the notion of credibility came up a lot. A consultant phrased this well:

> "The ECB needed to show that they really knew what was going on in the banks. They needed to take a deep dive. And they needed to set common ground rules for all the banks. We saw this in the early CEBS exercise. It was worth nothing, and nobody believed it. Banks were still all doing whatever they wanted with all the national discretions. It said the banks were all fine, and then after publication we saw banks failing. That's why the ECB needed this elaborate rule book for banks to follow."

This shows that in order to be a credible indicator, commensuration plays an important role. Supervisors cannot just produce a list of numbers, and hope that the wider public will believe them. These numbers need to be made credible through processes of commensuration.

It is important to note that mobilising knowledge into the world is also a form of power. It reflects a choice in what will be made visible, and what is left invisible. Besides revealing the overall results of the stress test, the EBA also publishes thousands of bank-by-bank data points on its website, in the name of full transparency. This idea of transparency has achieved cult status. It is hyped as the panacea to backdoor politics and corrupt politicians. However, making too much knowledge publicly available can have a destabilising effect too. Although more information can result in an empowerment of society, it can also lead to the exact opposite. A chief risk officer in a bank noted:

> "Have you heard of the saying too much tax kills tax? Well it also applies for information. Too much information kills information. At first, I was very cautious about all of these data points going

> public, I was worried. I expected a lot of phone calls or reactions. But, there was not much. No one really takes the time you know. What we do is, we make sure that when everything gets published by the EBA, we have our story ready. We can explain the results and the data and everything ourselves. And that is what people look at in the end."

Although the EBA makes thousands of data points available, information consumers still predominantly rely on the reports and summaries published by the EBA or the banks themselves, and see the information through the lens of these knowledge-producers. This shows that we need to remain cautious and critical regarding the role of transparency in democratising knowledge and empowering citizens. Where transparency might in some cases lead to a decentralisation of power, in others it can covertly strengthen existing power relations and dominant narratives.

How indicators measure: It's not about the numbers

The production and use of policy indicators in global governance is increasing rapidly. Indicators have the unique power to simplify policy issues and rank performances according to a simple numerical scale. As such, they have important regulatory effects, even when they lack any legal mandate. In order to make a regulatory indicator, different (national) systems need to be commensurated, i.e. be made comparable according to a common metric. Despite a burgeoning literature, little empirical work has been done to further our understanding of the social and political processes through which these

indicators measure and make units measurable (Davis et al., 2012; Huault & Rainelli-Weiss, 2011; Peeters & Verschraegen, 3013). Where it is often assumed that regulators have a blind preference for commensuration, and standardisation at any cost (Gorur, 2016; Scott, 1998), this chapter shows how regulators allow 'incommensurable' categories to exist due to largely unrecognised regulatory benefits.

This work draws on and contributes to the analytic tradition of quantification and governance by numbers, as developed by Desrosières (1998), Power (2003), and Rose (1991), and more closely to literature on commensuration (Espeland & Stevens, 1998; Kolk, Levy, & Pinkse, 2008). At the same time, it is located in relation to STS accounts, such as ANT and Callons' sociology of translations (Bijker & Law, 1992; Callon, 1984; Latour, 1987; Latour & Woolgar, 1979). What I add to this literature is an in-depth empirical understanding of how commensuration is done in practice, and why, giving us a better understanding of where the numbers we use come from. The bulk of the literature so far focusses on either how commensuration can be achieved technically, or why it is problematic and should be avoided. In this paper I use Callon's (1984) framework of translations to unpack commensuration as a multi-faceted process. This provides a more detailed notion of the motives underlying the commensuration process, as well as its benefits and drawbacks.

For starters, the results of the ECB stress test were seen as 'mild' and biased in favour of banks (Cecchetti & Schoenholtz, 2016). In both public and academic debate, regulation is seen to be routinely 'captured' and manipulated to serve the interest of regulated entities (Dal Bo, 2006; Etzioni,

2009). However, I find that regulatory capture proves to be difficult where interests of the regulated sector lie so far apart, and the sector sees itself as incommensurable. Additionally, I find that although regulators could have made a more 'game-proof' indicator, their choice to abstain from doing so did not result from industry pressure but was a conscious choice. This adds an important nuance to theories that use regulatory capture to explain regulatory outcomes and designs (such as Bunea, 2013; Hanegraaff et al., 2016; Klüver, 2013; Lowery, 2013; Woll, 2009).

Secondly, I challenge the idea that some things 'are' more commensurable than others. Banks across Europe are not automatically comparable, they need to be made so. This is often overlooked in policy fields like the financial sector, where quantification is taken for granted. Our findings emphasise that commensurating systems is a socio-political construction that requires a lot of effort. In this vein, I argue claims of incommensurables can be made, and should be heard, in all policy fields. Allowing political actors to make claims of incommensurability, is a way to emancipate them and give them a critical voice. As such this argument ties into critiques of instrumental reason (for instance, see Smulewicz-Zucker, 2017).

Moreover, I find that commensuration can contribute to credibility. A large part of the credibility of the stress test results from the extensive commensuration efforts, as proof that the ECB is assessing European banks on a Level Playing Field, subjecting them to extensive uniform rules. As a key policy indicator, it was imperative that the stress test was taken seriously by market participants. If the ECB would have published a list of banks' health

that was not taken seriously by the wider public, this could have aggravated the crisis.

Finally, however, I find that banks are still treated as incommensurable to some extent. The ECB does not assess banks' assets according to standardised models. Rather, banks are allowed to use internal models for calculation. This may seem as if regulators are giving banks free play to game the indicator, and make their results look as good as possible. However, this choice for incommensurability is informed by other regulatory benefits, such as shared accountability, information access, and improved internal risk management capacity. Although literature usually promotes the design of game-proof indicators (Bevan & Hood, 2006; Hood & Peters, 2004; Politt & Talbot, 2004; Smith, 1995), I argue that allowing room for interpretation regarding the various measurements in an indiator can have regulatory benefits that have largely gone unrecognised.

This ties into Barry's (2012) notion of knowledge controversies in transnational governance. Regulators shy away from establishing an extensive transnational consensus on how risk should be understood and calculated. On the one hand, they are apprehensive of bearing the responsibility of establishing a standard model of measuring risk. On the other, they seem to believe that knowledge controversies, to a certain extent, allow for innovation, as they maintain a continuous reconsideration of how to understand and measure risk. This goes to show that technological instruments, such as the stress test, play a critical part in developing regulatory spaces (as also argued in Barry, 2001). This non-instrumentalised conception of knowledge is also reminiscent of Habermas' (1984) notion of

communicative rationality as an alternative to instrumental rationality. The former focusses more on increasing understanding through open communication, while the latter is more strategic and results-oriented. Processes of commensuration tie in closely with instrumental rationality; as they both aim to manipulate the world in order to control it. Following this, indicators can act as a standardising instrument for control and domination (Marcuse, 1941). Treating banks as incommensurables then reconceives knowledge making as an ongoing exchange among critical equals, rather than a fixed outcome of a so-called rational process imposed by dominant actors. Allowing incommensurability establishes discursive conditions that offer a more critical and understanding-oriented space for the regulatory exchange. Rather than trying to merely measure and control risk by imposing 'rational' knowledge, regulators here make a more critical attempt to understand risk. Recognising banks as incommensurables thus marks a noticeable shift from an instrumental to a more critical epistemology in financial regulation.

Are stress tests breaking the bank?

This chapter addresses the question of whether the stress test is breaking the bank, as a question of whether stress tests (and indicators in general) are tough enough. Indicators often receive criticism of being biased, because it is said that they are captured or gamed by industry interests. So too, the stress tests have not been exempt from criticism, mainly regarding the severity and credibility of the results (Dowd, 2015). It is fair to say that the macro-economic scenario was less stressful than its US counterpart (Enria 2018).

However, it is important to qualify that this was not because European regulators were captured by industry interests as critics are eager to claim. This chapter clarified the role of bias in measurement, and how bias can be used strategically. Allowing banks some leeway in calculating the stress test stimulated discussion and deliberation between regulators and regulatees regarding how risk should be measured and understood. Assuming that banks, or other entities, can be easily measured according so some so-called rational logic, is problematic, especially in high-risk environments. Instead, allowing incommensurability in the measurement process establishes discursive conditions that offer a more critical and understanding-oriented space for the regulatory exchange.

Overall, I would argue that, although the scenario of the stress test might not have been as tough as in the US, the stress test was a severe exercise for banks. Although the bottom-up approach may have left more room for gaming efforts, it also created space for learning opportunities.

How indicators manage:

Using numbers that don't count

This chapter is published in a slightly adapted form as:

Kempeneer, S. & Van Dooren, W. (2019). Using numbers that don't count: How the latent functions of performance indicators explain their success. *International Review of Administrative Sciences*

Chapter overview

This chapter asks under which conditions performance indicators can improve performance outcomes. I show that performance measures do not only manage by results, but already manage behaviour latently through the design of the measurement process. In the case of the stress test, banks are not likely to change their behaviour based on the results of the indicator; they even find this information to be invalid. Instead, through the process of having to calculate the exercise, banks configure new habits and patterns of behaviour, encouraging processes of self-regulation, and inadvertently improving performance outcomes. Moreover, I show that ritualistic functions of indicators can improve actual performance outcomes, and invalid measurement does not necessarily hamper performance outcomes.

THE PERFORMANCE MANAGEMENT PARADOX

Since the global financial crisis, stress testing has become part and parcel of regulators' toolkits for monitoring and maintaining financial stability globally. At first sight, the success of stress tests, is easy to explain. As awareness for risks increases, the demand for instruments that measure and control risks is growing (Power, 1997). Performance indicators' success hinges on their ability to simplify, standardize, compare and control complex systems. The EU-wide stress test delivers on this demand for control. The stress tests are designed to show at a glance the banks that would weather out a crisis.

However, decades of research have also exposed the adverse effects of performance indicators just like the stress test. These indicators are said to lack accuracy, encourage gaming, demotivate workers, be biased towards what is quantifiable, and ultimately fail to substantially improve performances (Berten & Leisering, 2017; Bevan & Hood, 2006; Bouckaert & Balk, 1991; Davis et al., 2012b; Hvidman & Andersen, 2014; Pollitt, 2018). Bevan & Hood (2006) even liken management by indicators to the failed Soviet planned economy in the 1930s and 1940s. There are worryingly common anecdotes of how measurements are distorted to create an illusion of good performance. Take school rankings for instance. Schools with well performing students get good rankings. An unfortunate consequence of these rankings is that teachers are asked to 'teach to the test', to only impress on students that which will be examined, in order to improve students' average grades and as such the overall ranking (Gorur, 2016). This creates the illusion that highly ranked schools have a high quality of teaching, causing students to succeed. While in reality, teachers are demotivated and students do not learn

anything beyond the end of term requirements. Research has also identified misuse of performance information as a key factor hampering performance improvement (Micheli & Pavlov, 2017; Moynihan & Kroll, 2016; Taylor, 2011; Van Dooren et al., 2015).

Some scholars even go so far as to claim 'evidence based' and 'rational' policy making is a myth, because policy work is fundamentally political (Boswell, 2018). Every measurement is seen as political, because there is always a (political) choice of what can or should be measured. For instance, the choice not to include unpaid work or environmental costs in GDP is a political choice. These adverse effects are primarily at play when highly incentivized performance indicators are used, when there is much at stake (Bevan & Hood, 2006; Van Dooren & Hoffmann, 2018). The stress tests have not been spared this criticism, especially in financial media. Stress tests are said to not make sense economically, be biased towards certain banks, and be little more than communication exercises to reassure financial markets (Cecchetti & Schoenholtz, 2016; Dowd, 2015; Elliott, 2016)).

To be sure, there are studies that show that performance measurement positively affects performance outcomes (Boyne & Chen, 2006; Nielsen, 2014; Walker, Damanpour, & Devece, 2011). Yet, we do not have a clear understanding of why performance indicators improve performance outcomes in some cases and fail to do so in others.

Despite these differing and seemingly contradictory effects of performance indicators, regulators continue to promote them as indispensable tools for regulation. Over the past decades the use of performance indicators has proliferated in (global) governance (Davis et al.,

2012b). The ambition of this chapter is thus to unravel this management paradox and explore under which conditions indicators can improve performance outcomes, despite their proven weaknesses and dysfunctions.

I find that performance indicators have important latent functions, that have so far been understudied in the literature. I borrow this concept from Merton's (1968) functional sociology. Merton distinguishes between manifest functions (intended and positive effects), dysfunctions (unintended and negative effects) and latent functions (unintended positive effects). Literature typically deals with the manifest functions and dysfunctions of performance information. In this chapter, I take a closer look at the latent functions.

First of all, I show that the process of calculating the stress test latently improves performance outcomes. To complete the stress test in a timely fashion, banks have professionalised their internal risk management systems; investing in enhanced data quality, improved IT-systems, and better coordination between risk domains. This inadvertently improves their performance outcomes. Secondly, I find that previously recognised latent functions, such as the ritualistic and symbolic function of performance indicators (Boswell, 2015; Power, 1997), are much more important for performance outcomes than they are often ascribed to be. Finally, I find that certain disfunctions, such as inaccuracy of results, do not necessarily hamper performance outcomes.

THE GOOD, THE BAD, AND THE UGLY

MANIFEST FUNCTIONS OF PERFORMANCE INDICATORS

Indicators abound in public governance. New Public Management (NPM) reforms in particular have led to the dissemination of indicators in all corners of government (Van Dooren et al., 2010). NPM aggregates a plethora of policy principles that all in some way intend to improve public sector performance by making it more efficient and goal-centred (Van Dooren et al., 2015), it does so through the principles of disaggregation, competition and incentivization (Dunleavy et al., 2005). For performance-based regulation to work, a key condition is that performance needs to be measured, and this is where performance indicators enter the picture. Many of the incentives that NPM promotes can only be applied when quantitative performance indicators are available.

Performance indicators easily found their way to the regulators toolbox as they provide information at a glance. They provide an objective measure of which organisations are reaching targets and who is underperforming. This information can then be used to improve performance outcomes by fostering competition between organisations, allocating resources according to performance, and increasing accountability (Braithwaite, 2014; Kagan, 1995; Levi-Faur, 2005). Moreover, the simplicity of performance indicators allows for more succinct communication with actors inside and outside government, tapping into an agenda of transparency and accountability; at least in theory (Sarfaty, 2011).

Osborne and Gaebler's (1992) seminal book 'reinventing government', clearly contrasts the advantages of governing-by-targets, with the inefficiency of bureaucratic governments spending tax-payer money as they please. It is worthwhile to quote the opening statement of the book, that nicely captures the atmosphere in which performance indicators came to flourish: "Are you disturbed and exasperated by the way government operates? If the answer is yes, and you seek to change the system, this book is for you." Throughout the book they continue this trend with motivational messages like 'what gets measured gets done', 'if you don't measure results, you can't tell success from failure', and 'if you can't reward success, you're probably rewarding failure'. All this to demonstrate the plentiful potential instilled in performance indicators.

DYSFUNCTIONS OF PERFORMANCE INDICATORS

Despite the abovementioned noble intentions, and high hopes, an increasing number of critical voices cite the paradoxical and dysfunctional effects of NPM and performance indicators, calling for a post-NPM reform (Christensen & Fan, 2016; Klenk & Reiter, 2019; Mikuła & Kaczmarek, 2019; Reiter & Klenk, 2018). After more than three decades of performance measurement in public policy, even sympathetic analysts, Like Hood & Peters (2004) or Dunleavy et al. (2005) acknowledge the adverse effect of NPM-reforms, especially performance indicators (Pires, 2011). Although in some (predominantly developing) countries NPM-reforms are still playing out, most advanced countries have come to realise that NPM has not fostered more effective or efficient public organisations. Instead, the NPM themes of disaggregation,

competition and incentivization led to siloed public bodies impeding collective action, perverse quasi-market mechanisms, and an obsession with intermediate organisational targets overshadowing service delivery and effectiveness (Dunleavy et al., 2005). In line with this, performance indicators specifically received a number of criticisms as well.

A first dysfunctional effect is that performance indicators may lead to tunnel vision. They tend to focus upon easily quantified dimensions of performance, thereby narrowing down the focus of policy-making and political debate to a small and often unrepresentative aspect of policy (Bevan & Hood, 2006; Pidd, 2005; Power, 1997; Termeer, Dewulf, Breeman, & Stiller, 2013). Doig, McIvor and Theobald (2006) add to this that an over-reliance on scores and rankings might overlook the fact that the phenomena they intend to depict are moving targets in terms of progress and direction. In the stress test, easily quantifiable risk areas such as credit and market risks have been addressed substantially, while areas that are more difficult to quantify, and difficult to pin down and define, such as operational risk are less developed.

Besides this, performance indicators can create perverse incentives and encourage 'gaming' and cheating. There are many empirical examples of how data is manipulated (Bevan & Hood, 2006; C. Hood & Peters, 2004; Pollitt & Talbot, 2004; Smith, 1995; D. A. Stone, 2002). For instance, hospitals will cancel appointments or schedule less follow up meetings to cut down waiting lists, creating an illusion of efficiency. Indicators are also said to stifle curiosity and diminish learning opportunities (Radin, 2006). Moreover, performance indicators may lead to goal displacement when organisations focus on the indicators rather than the underlying objective the indicators are

supposed to measure (Bohte & Meier, 2000). Furthermore, as O'Neill (2002) and Power (1997) have shown in their work on audits, performance indicators often obscure what is actually happening in the workplace, fuelling suspicion and mistrust, undermining professional ethics and generating a host of unforeseen problems.

Finally, the information performance indicators produce is often not even used or applied in decision-making (Johnston, 2004; Mol & De Kruijf, 2004; Pollitt & Talbot, 2004; Taylor, 2011; Walshe, Harvey, & Jas, 2010). Frequent causes are insufficient quality of the performance information, lack of important data, but also cultural or institutional barriers (Hoogenboezem, 2004; Van Dooren et al., 2015). De Vries (2010) adds that performance measures are usually a-contextual and unable to reveal anything substantive about the quality of politics. Performance indicators are just more red tape and paper work, wasting away in binders and computer folders.

This leaves us with a puzzle: with so many dysfunctions from research and practice being documented, under which circumstances can performance indicators actually improve performance outcomes? I claim that the answer is to be found in the latent functions of performance indicators.

THE LATENT FUNCTIONS OF THE EU-WIDE STRESS TEST

My interviews brought forward three key understandings of how performance indicators affect performance outcomes. I present my findings and theoretical interpretation simultaneously as to allow the reader to follow the abductive analytical process. First, I show how a common dysfunction of performance

indicators, inaccurate measurement, does not hamper performance outcomes. Secondly, I corroborate and complement the existing literature on latent ritualistic functions of indicators; showing that these can importantly affect performance outcomes. Finally, I show how the process of calculating the performance indicator can have a larger impact on performance outcomes than (the use of) the performance information itself. In calculating the stress test, banks made internal changes that improved long-term performance outcomes.

THE NUMBERS AREN'T RIGHT

A common dysfunction, addressed earlier in this chapter, is that performance indicators ultimately do not provide accurate performance information. Results are often said to be inaccurate, biased or gamed. In this section we examine how the actors involved perceive and deal with this apparent dysfunction.

Banks have unique assets in their portfolio that justify a unique way to calculate the risk weight of those assets. However, when given too much freedom, banks would end up with different risk weights even for very similar assets, gaming the system to their advantage. As such, the stress test methodology introduced caps and floors to somewhat level out the differences between banks' internal models. However, this common methodology was said to stand in the way of accurately reflecting banks' individual risk; raising questions about using the stress test to assess banks' performance. When we mentioned the EBA's common methodology to stress

testing teams, respondents sighed and started to shake their heads. A lot of bottled up frustrations flowed freely, as a respondent noted:

> "What you see in the EBA stress test is that you are put in a corset in terms of methodology. This is necessary to be able to compare banks, but it does not make sense economically."

Although the stress test makes a good effort of treating banks' risks and assets equally through the common methodology, the exercise sometimes lumps very different things together at the cost of accuracy. This supports the critical voices. The stress test might not paint a very accurate picture of each banks' actual performance, which might lead to unjust performance evaluation.

However, some nuance is required here. A goal of the European stress test was to do away with national bias and establish a European Level Playing Field (LPF). Although the stress test compromises on accuracy, without the LPF the stress test would, most likely, not be taken seriously at all. As many respondents pointed out, the early (2009, 2010) stress tests - where the common methodology was only a few pages long- gave banks substantial discretion in their calculations. Which, was often used to game results to banks' advantage. While today, the stress testing exercise is frustrating to banks (that are above all concerned with having an accurate result for their bank), all respondents agreed that overall the results paint a fairer picture of banks performance vis-à-vis each other. As such, I find that both regulators and regulatees agree that the stress test is dysfunctional, in the sense that it does not provide a completely accurate calculation of banks' performance

under risk, but it does minimize gaming, which leads to an overall better assessment of banks' performance.

Striking the right balance between providing a result that reflects the unique position of each bank, and confining gaming efforts and providing an LPF is difficult. In this case the stress test uses a bottom-up, rather than a top-down exercise. The previous chapter helped us understand this choice. A bottom-up exercise forces banks to develop their own models, and stimulates learning. At the same time, methodological constraints are put in place to curb gaming efforts. To some extent, this distorts banks' results and makes them 'less accurate'. However, it does create an LPF where all banks are subjected to the same constraints, and there are less opportunities for banks to window-dress their results. Although this balance might still not be optimal, it clearly holds many advantages.

RITUALS OF VERIFICATION

Though risk teams in banks were sympathetic towards the detailed rule-book and the LPF, they remained particularly frustrated about the granularity and intensity of the exercise. A respondent in a bank commented somewhat jokingly:

> "Risks are very specific, your clients can be pharmacists, and pharmacists are not butchers, it's a specific market, so the model needs to be specific. People with car loans in [one region], that's different from loans in [another region]. And each model depends on the behaviour of your clients, so you need behavioural

> parameters. That's what the internal models are for. And then what does the ECB do? They just add a buffer. But it's the same everywhere I guess. Engineers do this too, they make complicated calculations about how much cement they need and it's like 2,3658987 and eventually they're told, let's just take four. Everything is four. Always extra buffers."

Banks were left frustrated, not seeing the point of collecting all this granular bank-specific data. This can be understood by using the work of Power and others (C. Boswell, 2008; Gorur, 2015; Kelley & Simmons, 2015; Mahmood, Weerakkody, & Chen, 2019; Power, 1997), who have described performance measures and information as playing a symbolic role: they are valued as a means of signalling order and control. In his work on audits, Power argues that they operate as 'rituals of verification', providing assurances where there are low levels of trust (Power 1997; 2003). Just signalling that banks have to collect all this granular data and compute elaborate models, substantiates the claim that supervisors are digging deep, and being thorough. The pan-European stress tests symbolized a shift from 'biased' and 'weak' national supervision, to 'impartial' and 'rigorous' European supervision.

Besides a message of rigor, the simplicity of the exercise also worked to its advantage. The stress test can basically be presented as a ranking of banks in a crisis situation, making it easy to explain and disseminate to a wider public. This raised awareness that European supervisors were 'taking control', they were measuring banks' health and setting clear capital goals[40]. This

[40] The 2014 stress test even included an official hurdle rate of a 5,5% capital ratio, making it even easier to assess which banks were passing and failing the stress test.

added value of the stress test was picked up by respondents as well. During an interview a respondent confessed:

> "It's a very visible exercise. It helps to explain to people what it is I do. They've heard about it, seen it in the news. It gets more attention from a wider public. This is not just in De Tijd [a financial Newspaper], it's even on Het Journaal [the daily evening news]."

The stress test is a very visible performance measure, that also is expected to contribute to awareness and trust from wider publics. One particularly important public is found in the financial markets. The reassurance of financial markets cannot be underestimated. The stress test also was expected to send a signal of trust to the markets. Recall the explanation of the ECB employee quoted in the previous chapter, explaining that bias and mild results had less to do with capture and more with supervisors' strategies to steer financial markets. As he summarised "you're stuck with what you can afford". He explained that without a clear European backstop, and national governments across the EU committing to more austerity, it would have been problematic to present results with extreme capital shortfalls. That would only worry financial markets more, and made them even less likely to help recapitalise banks.

The performative character of the stress test is important in this regard. By saying that banks were doing fine, markets treated banks as such[41], buying the banks time to deal with their problems. Signalling trust in an

[41] Market credibility, and with it confidence, shifted in Europe after the ECB took over the stress tests (previously ran by the CEBS). As Anderson (2016, p. 9) writes in his comprehensive assessment "the tests were viewed as informative and credible. Equity prices and credit default swaps spreads moved substantially – improving for banks that were found to be healthy".

organisation can be key to allowing that organisation to improve their performance outcomes.

To be sure, just saying banks are healthy in the stress test is not enough, it needs to be a credible statement. In the early 2009 exercises banks scored well and faltered shortly after. In order to remain a credible exercise, regulators had to make sure that behind the scenes banks were cleaning up shop. The stress testing exercises conducted by the ECB and EBA contributed to this in a rather unexpected way. I elaborate on this in the next section.

More than just a ritual: Governmentality

As mentioned, banks are required to fill out extensive templates with over twenty thousand granular data points, over several risk categories. To do so, banks need to access granular data from all subsidiary branches in a short amount of time, be able to reconcile data from different risk departments, and explain in excruciating detail how various macro-economic variables will affect their assets. This did not merely serve the ritualistic or symbolic purposes stated above. Rather, it also, and more importantly, encouraged banks to improve their self-regulation. A respondent in a bank explained for instance how CEO's approved higher budgets for risk departments to improve their IT-systems, in order to successfully complete the stress test. These improvements in the IT systems are then used beyond the stress test, to improve banks' day-to-day risk management. Better IT systems help banks complete the stress test faster, but they also help banks detect problems and risks faster in their day-to-day business. Another improvement in this line, is

that the stress test brought people together over different departments. A stress test coordinator in a bank said:

> "It's a good experience to have, also for our internal stress tests. Because it's so intensive, you really need to go over everything, line by line. And you're also sitting at the table with so many people. That is also very important, this interaction between the different groups. Because when we do internal stress tests, it's not as thorough, and we're not sitting at the table with so many people. Here it's an important, and rich exchange of thoughts and methods, that is very valuable to think about stress testing in general."

This testimony shows how the EU wide stress test facilitated communication between different risk departments, as well as between risk workers and frontline workers in banks. These communication lines remained after the stress test was completed, again improving banks' internal risk management. This then improves overall risk management and performance outcomes.

Theoretically, I tie this to Foucault's notion of governmentality (Foucault, 2011); used to describe power that is exercised, not by directly regulating behaviour, but by steering how individuals or organisations self-regulate. In this concept Foucault brings together the notion of governing (gouverner) with modes of thought (mentalité). The government does not explicitly act upon an organistion, but the organistion acts upon itself. As such the term is often described as the 'conduct of conduct', the state-steering of self-regulation (Lemke, 2011). This emphasis on self-regulation can be seen as characteristic to the transition from liberalism to neoliberalism, as Renou

(2017) observes that the apparent withdrawal of the state actually marks a new kind of interventionism. Individuals and organisations are encouraged to take responsibility for themselves. Performance indicators typically act as tools of governmentality manifestly, by using performance information to make certain outcomes desirable (as demonstrated in Renou's work). However, we additionally find that performance indicators act as tools of governmentality by making certain practices and behaviours desirable and even necessary. While organisations can often readily game outcomes, gaming actual behaviour is much more of a challenge.

To be sure, the latent power exercised by the EBA and ECB is not against the interests of banks. Foucault is adamant that coercion is not necessarily bad. Moreover, this coercion does not mean that organisations are stripped from all their liberties to act as brainwashed 'puppets'. On the contrary, power, as it is discussed by Foucault, can result in an 'empowerment' or 'responsibilisation' of subjects with agency capacities (Bevir, 2010; Lemke, 2011). This empowerment and 'responsibilisation' is noted in the stress test as well. Supervisors do not simply hand banks knowledge about what is healthy or risky. The EBA's common methodology includes predefined categories, but banks still have the room to object or disagree (at least in theory). They are encouraged to think for themselves. Banks picked up on the learning experience they had through the stress test[42]. As a risk director noted:

[42] To be sure, large parts of conducting the stress test are seen as 'ticking boxes'. This has a lot to do with the restrictions in the common methodology. Risk experts in banks frequently vented their frustration with explaining meticulously how certain macro-economic events would affect their assets, only to receive yet another red flag because the end value exceeded a cap or floor.

> "We read papers and we follow workshops and we try to keep up, but it's not always easy to find time. The stress test forced us to look at things that we had been neglecting. So, it's a good learning experience. It's a new kind of learning, not just from books, but learning as you go."

At first sight, it seems that the stress test just assesses banks' performance in a stress scenario, and penalizes accordingly. Banks that see a big drop in their capital ratios, need to recapitalize or de-risk their portfolio. However, there is more going on. The stress test does not only passively map which banks perform well, but it actively shapes this performance. The stress test is a tool of governmentality, that coerces banks to act upon themselves, and improve performance outcomes, merely by calculating the performance indicator. The stress test is thus latently constitutive of the performances it measures.

How indicators manage: Latently

Over the past decade, stress testing has become part and parcel of banking regulation in the EU. The EU-wide stress test calculates how banks would fare in a hypothetical plausible yet adverse stress scenario. Many comparable indicator regimes in (inter)national governance have been criticised for generating dysfunctional effects.

This chapter addressed under which conditions performance indicators, such as the stress test, can improve performance outcomes, despite their proven weaknesses and dysfunctions. The manifest objective of performance indicators is to measure performance, as to use this information

for learning, steering and control, or accountability (Van Dooren et al., 2015). Besides these manifest functions and the often-accompanying dysfunctions, I argue that indicators fulfil important latent functions as well; that have so far been largely overlooked. Based on interviews with stress testing teams in Belgian banks, the National Bank of Belgium, consultants, and officials at the European Banking Authority (EBA) and European Central Bank (ECB) this chapter took a closer look at the latent functions of performance indicators, and how they can contribute to performance outcomes.

First of all, I found that what is commonly seen as a dysfunction of a performance indicator, need negatively affect performance outcomes. Like many other indicators, stress tests seem to face validity issues. These seem to challenge the manifest goals of performance indicators, i.e. accurately measuring performance in order to steer performance outcomes (Van Dooren et al., 2015). However, dysfunctions such as inaccurate measurement can serve an important role in furthering the overall objective of improving performance outcomes. Compromising on accuracy proved to be a necessary part of a trade-off to ensure a level playing field, which was key to the overall credibility and legitimacy of the stress test, allowing it to improve performance outcomes.

Secondly, my fieldwork corroborates and complements earlier findings in literature, that indicators fulfil important ritualistic functions (Boswell, 2008, 2015; Power, 1997). They can signal that governments are dealing with a problem thoroughly, and that accountability mechanisms are in place, instilling trust. In this chapter I show that these latent functions can also be key to actually improving performance outcomes; a quality that is

often overlooked. By stating that an organisation is performing well, and that regulators are on top of the situation, organisations are given the necessary room to actually work on improving performance outcomes. This mechanism is known as performativity (MacKenzie, 2006). As such, these ritualistic functions of performance indicators should not be brushed off as a pleasant side-effect of performance indicators, rather they should be more widely recognised as key factors in allowing organisations to improve performance outcomes.

Finally, I show that the process of calculating a performance indicator can latently improve performance outcomes in itself. Calculating the stress test inadvertently caused banks to professionalise their risk departments, which improved banks' internal risk management, in its turn improving long-term performance outcomes. I explain this mechanism drawing on Foucault's theory of governmentality (Foucault, Burchell, Gordon, & Miller, 1991): Regulators improved performance outcomes by operating on a latent level; by educating and configuring habits, aspirations, and beliefs. In the process of calculating the performance indicator, banks updated IT-systems, increased communication across departments, and improved internal processes. Although they initially only made these changes as part of the process of calculating the performance indicator, they ended up actually improving their internal risk management, and thus their performance outcomes. Banks' performance outcomes are not improved because they learned from the performance information itself; on the contrary, they even find the performance information to be invalid. Rather, they improved their performance outcomes by revising internal management systems to be able

to calculate the performance information. This mechanism has been largely overlooked in literature so far, warranting more scholarly attention. The process of how the performance indicator is calculated might be a key factor in explaining why performance indicators succeed in some cases and fail in others.

To conclude, the manifest goal of performance indicators is to produce performance information that can be used to improve performance outcomes, by allowing organisations to learn from or reflect on this information, or by rewarding and penalising over- and underperformers. However, an increasing number of critical voices show that in many cases performance indicators fail to improve performance outcomes, because of disfunctions such as gaming or manipulation for political power-plays (Davis et al., 2012b; Dunleavy et al., 2005; C. Hood & Peters, 2004; Van Dooren et al., 2015; van Thiel & Leeuw, 2002). The main contribution of this chapter is that it shows new, latent, ways in which performance indicators affect performance outcomes. I show how the process of calculating performance indicators, which is often overlooked in literature, can in itself improve performance outcomes - regardless of the results of the indicator, and their use or validity. Where a key criticism of performance indicators is that performance information is often inaccurate, biased or invalid, I thus rebut with the afterthought that using numbers that don't count can latently help to manage performance outcomes.

Are stress tests breaking the bank?

This chapter addresses the question of whether the stress test is 'breaking the bank', as a question of whether the benefits of the stress tests are worth the costs, or if it is just a waste of resources? I would argue that in order to weigh the costs and benefits, one needs to consider the wide array of benefits that span beyond the scope of the stress test. The stress test is without a doubt a costly exercise. Respondents in banks found it especially frustrating that they were required to invest their resources in providing granular information, where in the end these did not seem to matter much, and methodological constraints did not allow for accurate results after all.

However, in order to access this granular data from subsidiary branches, banks were encouraged to optimise their internal risk management systems. For instance, they reconciliated data across diverse systems, built IT-tools to automate large parts of the stress testing exercise, and created more formalized communication structures across risk departments. These changes were only possible because CEO's were more inclined to invest in risk departments with their reputation at stake in the stress test. The changes did not only help banks complete the stress testing exercise more efficiently, but they also continued to prove their use in banks' day-to-day risk management. Moreover, as the regulators further optimise the exercise, it is likely that the costs will be reduced.

Overall, I would argue that the stress testing exercise has been worth the cost, given the transformations it has instigated in the banking landscape.

How indicators make: A Big Data State of Mind

Chapter overview

In 'describing' the world, indicators have the ability to (re)make the way we see the world, and ultimately affect what we accept as truth. In this chapter I explore how using indicators based on large data sets is accompanied by a 'Big Data State of Mind'; the epistemological notion that one can or should rely on large data sets rather than theory to observe and understand reality. I look at this state of mind in the EU-wide banking stress test. Through interviews with respondents in Belgian banks, consultants, and supervisors at the European Central Bank (ECB), European Banking Authority (EBA), and National Bank of Belgium (NBB), I explain how the shift to a Big Data State of Mind has an important impact on the behaviour of regulators, and their relationships with regulated entities. I especially discuss the inherent problematic relationship between a Big Data State of Mind and traditional notions of accountability and transparency. I advocate a transparency based on dialogue rather than disclosure.

MAKING DECISIONS WITH DATA

There is a strong celebratory thread in the literature on Big Data; that more data will bring better science, safer cities, healthier citizens, and rapid innovation. One such book is 'The Human Face of Big Data' (Smolan & Erwitt, 2012), a collection of essays about the potential of Big Data to design personalized drugs, predict divorce, and research Parkinson's disease. In the public sector too, Big Data promises to improve decision-making and the overall effectiveness of government and regulation (Desouza & Jacob, 2017; O'Malley, 2014; van der Voort, Klievink, Arnaboldi, & Meijer, 2019).

As the second chapter showed, governments have a long history of relying on data, and increasingly do so in various aspects of their functioning (Hazen, Boone, Ezell, & Jones-Farmer, 2014; Janssen & Kuk, 2016; Matheus, Janssen, & Maheshwari, 2018; Mayer-Schönberger & Cukier, 2013). One application is in the development of policy indicators (Niemeijer, 2002). Policy indicators are important tools in governance, as they succeed in presenting a standardised and simplified view on reality. Indicators can track individual performance, public sector performance, as well as performances of regulated sectors. This is important for decision-making (for instance when indicators are used for funding allocation), but also to communicate and interact with wider publics. In their work on dashboards Matheus, Janssen and Maheshwari (2018) describe how data visualisation allows publics to scrutinize government actions and engage in decision making. The same dynamics apply to indicators as they also allow publics to easily keep track of performance outcomes and government interventions.

Niemeijer (2002) already suggests the difference between theory-driven and data-driven indicators. He states that for the data-driven approach, data-availability drives what will be included in the measurement, whereas for the theoretical approach the focus is on selecting specific data from a theoretical point of view.

For theory-driven indicators, good performance in a specific domain is theoretically connected to a set of factors. However, as this paper will show, it is not always straightforward to theorize what good performance looks like, and which variables should be included. It is important to note that when we build a theoretically informed model of the world, or a theory-driven indicator, there is a limit to the number of variables we can consider (the number of independent variables you can include is limited by the sample size). So, we make a clear theoretical choice to include some factors and exclude others. For instance, the OECD's Programme for International Student Assessment (PISA), often used in educational reform, covers reading, mathematic, and science skills, but leaves out other competences such as art or language skills (OECD, 2019). It is a theory-driven indicator. To be sure, as later parts of this chapter will show, it is not always straightforward to choose which variables to include and what to leave out.

For data-driven indicators, those choices don't have to be made; the objective is to look at anything and everything available. In an era of Big Data, more and more data are (automatically) gathered, often continuously. The idea is that relying on large amounts of data would lead to a more accurate representation and observation of the world (Anderson 2008). Following this train of thought, Peter Norvig, Google's research director, is quoted stating

"All models are wrong, and increasingly you can succeed without them." (Anderson 2008). This idea is well summarized in Chris Anderson's (2008) provocative and widely cited blogpost 'The End of Theory'[43], where he states:

> "Theory is dead, long live data! (…) Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."

This quote remarkably captures the epistemological component of Big Data, that I call a 'Big Data State of Mind'. Rather than focussing on the concept of Big Data as such, this chapter draws particular attention to the shift in mindset that accompanies the use of large and often unstructured data sets. I take the case of the EU-wide banking stress test to explore to which extent this Big Data State of Mind is present, and how it affects regulation. To do so, I draw on interviews with risk experts in Belgian banks, national and European supervisors at the NBB, EBA and ECB, as well as consultants. I especially build on and contribute to the body of literature that analyses the relationship between Big Data and accountability (for instance Janssen & Kuk (2016), Kemper & Kolkman (2018), Vedder & Naudts (2017)).

[43] It should be mentioned though, that apparently Anderson never believed or advocated the theses of his own paper but wrote them to provoke response (see Norvig, 2008).

BIG DATA AND A BIG DATA STATE OF MIND

The previous section posited that more data, 'Big Data', might lead to better (policy) decisions. But what is big data? 'Big Data' is, in many ways, a poor term. First of all, in an era of unprecedented technological advancements, 'bigness' is a relative term. The 'big' data sets that used to require supercomputers, can now be analysed on any run of the mill desktop computer with standard software (Manovich, 2011). The size of the data is only one of many characteristics typically ascribed to Big Data. Despite a lack of consensus on what Big Data is precisely, most authors agree on the "Three Vs"[44]: Volume, Variety and Velocity (Chan & Moses, 2016; Kitchin, 2014; Salganik, 2017). Besides being 'big' in volume, the data typically comes in a variety of formats and is being created constantly at a high velocity.

But, Big Data is about more than the three V's. Big Data carries an epistemological component as well. Epistemology is the theory of knowledge, it relates to what knowledge we accept as true and how we learn about the world. Both ways of observing the world, theory-driven and data-driven, represent a different epistemology. As for instance Boyd & Crawford (2012, p. 123) argue, Big Data "reframes key questions about the constitution of knowledge, the process of research, how we should engage with information and the nature and categorization of reality." In sum, Big Data promulgates the idea that in order to understand the world, we no longer need theories about the world, just more data.

I call the epistemology that relies on Big Data, a 'Big Data State of Mind'. I use the label of a 'state of mind', because the discrepancy between

[44] Critics with a sense of humour commonly add a V for Vague as well.

a reliance on theory and a reliance on (big) data forms a continuum. For instance, this ranges from relying on the presumed theoretical inverse relationship between inflation and unemployment to develop monetary policy[45], over using an algorithm to make a first selection of potential hazardous traffic situations, but prioritizing funding based on theoretical models, to insurance companies relying solely on algorithms to determine premiums. It is the shift in mindset (regardless of the stage of full reliance on data) that is of particular interest in this chapter. I am not strictly concerned with the completed practice of solely relying on big data and automatic algorithms, but the state of mind that one can or should rely on large data sets rather than theory to observe, understand, and control reality.

There are three key characteristics of this epistemological mindset (Mayer-Schönberger & Cukier, 2013). First, the idea of *comprehensiveness*; the idea that large amounts of data will provide a comprehensive perspective of the characteristics of a phenomenon. This ties in with the idea that we no longer need theory to know or choose where to look. Instead, we can look at everything. The second is the idea of *messiness*. The world is multifaceted and complex and no theory will ever be able to contain or reflect this complexity. As such, the only way to study this messiness is by gathering (messy) Big Data that will automatically reveal existing patterns and correlations. The third idea, the *triumph of correlations*, ties in with this: The idea, also voiced by Anderson (2008), that we no longer can (or need to) understand why patterns occur (causation). In the words of Mayer-Schönberger and Cukier (2013:14) "The correlations may not tell us

[45] Which interestingly has been proven to only hold true on the short-term, since more data has become available since the 1970's.

precisely why something is happening, but they alert us that it is happening. And in many situations, this is good enough." Moreover, often with one click of the button, big data techniques can instantly determine which model has the best fit (without researchers engaging in p-hacking, and iteratively adding and removing variables). This mindset would lead us to believe that a data-driven representation of reality would be a more accurate reflection of reality, than a theory-driven one.

At first sight, the idea to gain insights directly from data does not seem like an entirely unfamiliar epistemology. There is a longstanding inductive[46] or explorative scientific tradition where raw observational data is used for theory development rather than hypothesis testing. However, this approach does not ring in the end of theory, as a key objective of this epistemic approach is to develop new theory (Stebbins, 2001). The premise of a Big Data State of Mind is that theory is no longer relevant at all, because as long as correlations can be identified, there is no need to understand the theoretical mechanisms that explain 'why' things are correlated (Mayer-Schönberger & Cukier, 2013). This is a more controversial and interesting thought; whether a large amount of data completely obviates the need for theory. This interpretation is a more extreme empiricist epistemological judgement of what kind of knowledge is useful.

[46] To be sure, the inductive approach has received its fair share of criticism too. As Karl Popper (1963: 123) wrote: *"the belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society to be used as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not."*

It is this epistemological shift to a Big Data State of Mind that lies at the base of much of the criticism geared at Big Data (Kitchin, 2014; Frické, 2013; Couper, 2013; Vigen, 2015; Symons & Alvarado 2016). A Big Data State of Mind, is considered to be a threat to the quality of knowledge, due to its complete disregard of theory. This is deemed problematic because correlation can just be a coincidence[47], theory is necessary to understand how two variables are connected and why. For instance, eating a lot of chocolate can be correlated with intelligence (as in Messerli's (2012) paper linking chocolate consumption to Nobel laureates), but because there is no theoretical explanation why chocolate would make anyone smarter, this is likely just a coincidence. Correlations without theory are of little use to scientists[48]

The disregard of theory also makes it increasingly difficult to understand where policy-decisions come from, and whether they are impartial and just. Cathy O'Neil (2016, p. 1) writes in her book 'weapons of math destruction' that "we live in the age of the algorithm". Important decisions that affect our lives are being made by mathematical models. On the one hand, this is supposed to lead to better judgement, because life is messy and complicated and humans and their theories are flawed and algorithms based on large data sets might to a much better job at describing the world for us (Anderson 2008). On the other hand, if an algorithm is trained

[47] Tyler Vygen's (2015) book 'Spurious Correlations' compiles dozens of coincidential correlations between completely unrelated sets of data. For instance, linking cheese consumption to bedsheet tangling accidents, or margarine consumption to the divorce rate in Maine.

[48] To be fair, most Big Data techniques account for spurious correlations through measures of interaction depth between variables.

on data that is biased, it may learn to continue to discriminate, reinforcing social inequalities and bias (Diakopoulos, 2016). As Janssen and Kuk (2016, p. 371) write in an editorial: "as algorithms become increasingly autonomous and invisible, they become harder for the public to detect and scrutinize their impartiality status". This shows how the epistemological shift towards a Big Data State of Mind has important political consequences. Algorithms often act as black boxes in decision-making, necessitating an important discussion regarding accountability.

Take the case of the use of the COMPAS algorithm in the US criminal justice system, described by Kirsten Martin (2018). Martin tells the anecdote of an exemplary inmate who, when brought to the parole board, was denied a transfer based on COMPAS that neither the inmate or the board necessarily understood or agreed with. The choice not to disclose more information regarding COMPAS is defended by the claim that it is a 'trade secret' – that cannot be disclosed to avoid gaming efforts. In other cases, transparency efforts are in competition with privacy rights.

Bambauer (2017) argues that full transparency might not be feasible nor desirable for precisely these reasons. However, several authors (for instance Burrell, 2016; Pasquale, 2015) warn that intentional obscurity can be designed to avoid scrutiny, which is equally harmful to the overall system. A balance in transparency needs to be found so supervisors can still be held accountable for decisions based on big data, whilst this transparency does not jeopardize regulatory scrutiny. This trade-off ties in with important discussions in public administration regarding conflicting sets of core values. Hood (1991) distinguishes three sets of core values: sigma-type values

(efficient governance), theta-type values (equitable and fair governance), and lambda-type values (resilient governance). Full transparency ties in with values of fair governance, however it can challenge government efficiency and resilience.

The question often boils down to how much transparency, or justification for a decision is enough? As Binns (2018) writes, full transparency could involve sharing lines of code, and intricate data matrixes that describe what is going on during the algorithmic transformations, but this would still not shed any light on why the algorithm is making these transformations. Binns (2018) suggests that instead of focusing on disclosing details of algorithms, decision-makers can justify their decision based on previous success of an algorithm, or by the scientific rigour involved in developing it, without fully disclosing (or themselves understanding) how the algorithm works exactly. This might be sufficient to trust the decision, but critics might also still remain sceptical. The question is thus an epistemic one; under which circumstances, are we willing to accept data-driven decisions (or data-driven knowledge), that we might not fully understand. Take the case of the COMPAS algorithm again. Perhaps the inmate only seemed exemplary to a jury, and the algorithm had good reason to deny that transfer based on the large base of information it draws on.

In sum, a Big Data State of Mind requires an epistemic shift that places its wager on large quantities of data to help us resolve societal problems. In what follows, I will show how the EU-wide banking stress test reflects this epistemic shift to a Big Data State of Mind, and evaluate how it affects

regulation. Related to this, I show how this State of Mind challenges processes of accountability in the stress test.

THE STRESS TEST AND A BIG DATA STATE OF MIND

WHEN THEORY FAILS, AN ARMS RACE FOR DATA

In a sense, the 2008 financial crisis was a crisis of theory. Regulators, banks, and financial markets all had encompassing models about how the economy worked, but they all failed to predict the looming crisis. The theoretical models had lost touch with the actual economic reality. A lot of the assets that banks held during the crisis got safe ratings, causing market participants to view them as safe. As such, credit rating agencies were pinned as one of the main culprits in the global financial crisis, for creating such a misleading depiction of banks' health (Rafailov, 2011). In the wake of the crisis, it dawned on supervisors that reality is chaotic, and despite all their efforts it would be a tall order to discern whether banks were healthy or not. This ties in with the idea of *messiness*, that the world is too complicated and messy for a theory to contain, which was mentioned earlier as one of the key drivers of a Big Data State of Mind (Mayer-Schönberger & Cukier, 2013).

This can be seen as somewhat of a turning point for financial supervision, carrying implications for risk management at large. The world around us has become so complex, that it might have grown beyond our comprehension and control, somewhat like Frankenstein's monster. As a society we have created a banking system that we now have lost our grip on.

During an interview at the ECB, I had an interesting conversation about the difficulty of predicting risk factors for large and complex banks:

> "Take asset pricing models, the idea is you take the share price of a firm and you base it on macro factors, and the ones that come out significant are the ones that affect the risk of your firm. (…) So, we did this for a few banks, and for this big bank, [bank X], nothing was significant. So, they don't look correlated to any of the macro factors. So, does that mean that [bank X] is risk free? Of course not! It means that the risks of [bank X] are such that they are very hard to pin down on something that we have a clear and easy grip on."

This quote shows how difficult it is to predict which factors will drive a good performance for a bank. This leaves supervisors in a tricky situation, because although the world is becoming increasingly difficult to model, political actors cannot simply refrain from making plans. A respondent at the National Bank elaborated on the new, more data-driven, approach in the stress test:

> "People start thinking 'oh no, how come this crisis happened'. Maybe some things went wrong in terms of supervision, or maybe the supervisor did not have the right data to see it coming. (…) And for every part everyone says 'well I want very granular detailed data now'. Because no one wants to miss it again. (…) and so, you end up with stacks and stacks of additional requirements. And it's hard to clean house, because no one knows where the next crisis is going to come from. I don't think

> we can have the arrogance to say that we know. You're always guessing."

This ties in with another driver of a Big Data State of Mind: *comprehensiveness*. The idea that large amounts of data will provide a comprehensive overview of the characteristics of a phenomenon (Mayer-Schönberger & Cukier, 2013). As a respondent in a bank added:

> "The reporting requirements [for the stress test] are mad. They increase every year. They ask for more and more granular data each time. We asked them in Frankfurt, when will it stop, this arms race for data. We're at 395 000 data points now, last time it was 260 000. And all they said was 'what's the alternative?'"

Because supervisors are no longer exactly sure which precise factors might drive a banks' performance, they increasingly rely on larger quantities of data; especially in the aftermath of the 2008 financial crisis. This trend of large data set capture and analysis by regulators is sometimes referred to as 'regulatory Big Data' (van Steen, 2015). Although this regulatory Big Data does not share all of the characteristics typically given to Big Data (it is not of the same magnitude as for instance a record of clicks on a popular website, nor is it always automatically generated), its volume is certainly much greater than the summary reports regulators typically request. This regulatory Big Data is expected to help regulators improve oversight and compliance, and enhance their understanding of the institutions they regulate (O'Halloran, Maskey, McAllister, Park, & Chen, 2015; van Steen, 2015). In what follows I examine how this Big Data State of Mind affects regulation.

In order to process the large amounts of submitted data, regulators rely on (semi-)automated big data tools and techniques (van Steen, 2015). One such tool used by the ECB is the Quality Assurance (QA) tool. As mentioned, the stress test is a bottom-up exercise where each bank calculates themselves how the crisis scenario would affect their banks. In order to make sure that banks are not 'gaming'[49] the exercise, all the data that banks submit is vetted by supervisors during the QA process. The setup of this QA process is pretty straight forward. The ECB uses all the data they have on banks to simulate themselves how the crisis scenario would affect the banks. To make these predictions, the ECB relies partially on Big Data techniques. If the results the ECB calculated are similar enough to what the banks submitted, it gets a green flag. If there is a minor discrepancy, it's flagged in orange. If there is a large difference between the results, there's a red flag. Banks' Joint Supervisory Teams (JSTs)[50] are in charge of discussing these flags with banks and resolving them. Banks can 'comply or explain', this means that either they accept the result that the ECB filed or they have to explain why their calculations were right after all, and should be left in[51]. A respondent in the NBB criticised this process:

[49] The notion of gaming was explained in earlier chapters, it refers to the idea that organizations can try to manipulate their data in order to achieve good scores on performance metrics (Bevan & Hood, 2006).

[50] JST's are comprised of employees of the ECB and representatives of the supervisory authorities of the member states in which the banking group operates. Each bank is assigned a JST as liason between the bank and the ECB.

[51] When there is a discrepancy between banks' results and supervisors' results, banks often de facto have to comply and accept the results that supervisors suggest based on their (data-driven) models. The stress test has tight deadlines, so banks do not always have the time to challenge the ECB's results.

> "I haven't spoken to all the JST reps, only some, but I feel like most of them are there to solve the flags but they're not actually trying to (hesitates) understand the results (laughs nervously). I was surprised - I'm not sure if I'm supposed to say this openly, but I was trying to understand the difference between the 2016 and 2018 results for some banks, and a lot of the JST reps had no idea why their bank was doing better or worse than in 2016. They're just focused on the flags. So, in a way I'm a little disappointed, if I can say that so openly, a little disappointed about this trend."

This came up during a conversation with a risk expert in a Belgian bank as well. The results of the stress test had just been published and the bank seemed to have done better than in the previous exercise. As I sat down with my respondent, I congratulated him on the good result. His reply was:

> "You know, it's funny, a guy from the JST also came to congratulate me on this. He really believed that we had de-risked our portfolio. But really, it was mainly due to a change in the methodology compared to last year. Nothing really changed you know. I'm laughing, but it worries me. They should know our bank."

Similar sentiments came up during many of my interviews. Supervisors seem overly focused on vetting the data and resolving the flags in the QA process, that they seem to not spend much time on understanding why a bank ends up with a given score on the stress test, which is very frustrating to banks. The stress test often acts like a black box that just predicts an outcome without anyone understanding, or more importantly justifying, where that outcome

comes from. This is in line with the epistemic positioning that correlations are deemed more important than causation, which was the third key element of a big data state of mind (Mayer-Schönberger & Cukier, 2013). To banks, supervisors seemed very uninterested in how they explained their results. When banks submit their data, they write accompanying 'narratives', explanatory notes that document why (according to them) their portfolios react to the stress scenario in any given way. One person said:

> "I'm not sure if the regulator has time to look at everything we deliver. (…) We didn't get any questions about our narrative. I don't know if they read it. We could have written anything in there I think. You spend all this time on that narrative, and you don't get any feedback, none. It's like you submit an essay in school and it's not graded. That makes you think, what have I been wasting my time on. I'm not sure they even need it."

Respondents in the NBB touched on this as well:

> "I feel like the people who write the explanatory notes, they must be so disappointed. Because they get flags and questions that they must think 'hey I anticipated this question in my explanatory note'. So, this [QA] tool that the ECB created, it's good for harmonization and uniform treatment, but it comes at the cost of a flexibility and informality where you can just look at banks' submission and ask questions about it, without all these flags. And it's not expected from the JSTs anymore. It's all about the QA tool. I just want to say to a lot of these JST reps: 'just call the bank

> and ask why you have this result', instead of this constant focus on the QA tool."

Respondents feel that supervisors often blindly accept and impose the conclusions brought forward by their data. If a discrepancy between banks' results and the ECB's results is flagged, banks are often de facto forced to accept the ECB's result without a clear explanation or justification of how this result was obtained and what it means.

BLACK BOXES, BIG DATA, AND ACCOUNTABILITY

This has an impact on processes of accountability. As a respondent in a bank complained:

> "Suddenly your results are replaced with something new, something that does not make sense to you. And when you ask for more information, they do not disclose anything. It is very unclear what models the supervisors are using. And that makes it difficult to try and make sense of this exercise. It makes it difficult to understand why they think a portfolio is at risk, or where certain losses supposedly come from. That's why I don't really mind the results of this exercise so much anymore. They are not very useful for me or my team internally."

This quote shows that the results obtained by ECB do not always make much sense to banks. As mentioned, the results of the stress test feed into banks Supervisory Review and Evaluation Process (SREP). But,

because the results of the stress test are opaque, and the way in which the stress test feeds into the SREP is equally opaque, banks are often dissatisfied with the entire process, asking for more transparency. A consultant I spoke to even added:

> "I've heard that banks have filed an appeal against the SREP. All I know, is that apparently in 2014 eight banks filed claims, and four won. I also heard, but again I just heard this, that there is a law firm in Frankfurt that specialises in this kind of thing now. And that's legitimate in my opinion. It can't be the case that the ECB is just randomly bullying banks with capital requests. They should have to justify their decisions. That's crucial for the integrity of the entire dialogue. (…) I heard that these four banks won because the SREP decision and the observations were not documented enough to support the decision. So, everyone is accountable, and must be held accountable for what he does."

This shows that the move towards a Big Data State of Mind, necessitates an important discussion regarding transparency and accountability. Can supervisors continue to leave parts of their models undisclosed, and expect supervised entities to just accept the results without a justification of where they come from?

The ECB (albeit justly) claims that their models cannot be disclosed to avoid gaming efforts. Respondents at the ECB noted that if they would reveal all the details of their models, banks might attempt to game the outcomes. Moreover, the ECB fears that if they disclose their models to banks, banks would no longer invest in developing their own models, which would be

detrimental to banks' internal risk management, and the safety of the system at large. This was touched on in chapter four as well, as an important disadvantage of full commensuration. Regulators wanted to avoid a 'mono-risk culture', where the ECB's models would become a unilateral guidance to banks. This clearly demonstrates the tension described earlier between values of equity (having fully transparent models) on the one hand, and efficiency (not risking any gaming) and resilience (the risk that banks would no longer invest in developing their own models) on the other. Diffusing this tension is a clear challenge, especially when so much is at stake. In the case of the stress test, the QA process is seen as especially problematic because not only do supervisors not disclose their models, they provide little to no justification of their results in any way.

Without fully disclosing their models, supervisors could still provide more information to justify why their models might improve decision-making, rather than unilaterally imposing their results. Drawing on insights from chapter four, it would prove more helpful to view transparency from a framework of communicative rationality (Habermas 1990); as a tool to increase understanding through communication. Supervisors and supervised entities should engage in a dialogue on which knowledge claims will be accepted as valid, and under which conditions. Supervisors now ask banks for extensive narratives that they seemingly ignore, neglecting to open up any lines of communication with banks. As a consequence, respondents complain that supervisors fail to provide sufficient justification for the results they obtain, leading not only to misunderstanding, but also to mistrust.

HOW INDICATORS MAKE: SEEING THE WORLD THROUGH DATA

Indicators are typically seen as measures that describe the world we live in. However, in this chapter I showed how indicators can also actively make the worlds they intend to measure. Rather than seeing the world as it "is", we see the world through indicators. In the case of the stress test, the results tell us, and supervisors, which banks are healthy and which are not.

I then showed that there is a rise in data-driven indicators, rather than theory-driven ones. Where indicators are typically based on theoretical assumptions of how the world can be simplified, we increasingly see that supervisors rely on large amounts of data to understand what is going on (van Steen, 2015). This relates to wider societal trends of modernity and hyper modernity, that I discussed in the second chapter of this dissertation. In the ongoing modern mission to make the world around us legible and control it, this epistemic shift marks a new way of knowing about the world and understanding it. I called this shift a 'Big Data State of Mind'. This state of mind is related to three key characteristics (Mayer-Schönberger & Cukier, 2013). Firstly, *messiness*, the idea that the world around us is too messy and complex to be represented by a theory. Secondly, *comprehensiveness*, the idea that so many elements are intertwined in society, and only large amounts of data will be able to provide a comprehensive picture. Finally, the *triumph of correlations*, the idea that we no longer can (or need to) understand why patterns occur (causation).

Interviews with respondents in Belgian banks, consultants, and supervisors at the ECB, EBA, and NBB showed that this Big Data State of Mind is present in the EU-wide banking stress test, and that it has important

regulatory implications. Respondents in banks found it problematic that supervisors were not able to explain why banks performed well or not on the stress test. Moreover, supervisors themselves complained of colleagues showing little interest in understanding why certain results were found, assuming that the numbers must be right, with little justification regarding how the results were obtained. Large data sets can create many opportunities for better regulation, but are not without risks. The apparent benefits of Big Data should be balanced against the potential large economic and social costs of misguided policy decisions and mutual distrust (Tissot, 2017).

Furthermore, a Big Data State of Mind challenges current notions of accountability and transparency (Martin, 2018). A case can be made that data-driven decisions are less subjected to (biased) human judgement, and they might do a better job at comprehensively describing the complex and messy societies that we live in (Anderson 2008). However, when important decisions are based on data-driven models, at least some quality assurance and justification should be provided for these knowledge claims. Supervisors at the ECB claim, in line with arguments made by for instance Bambauer (2017), that transparency regarding their models is not feasible nor desirable because it enables gaming efforts which could be destructive to the regulatory system. On the other hand, intentional obscurity is equally harmful. Take for instance the banks filing complaints against their SREP decisions because the process is opaque and they feel the capital decisions are unfair. This means we need to rethink what kind of transparency is useful and necessary in the 'age of the algorithm'.

A good way forward might be to move towards a transparency that leads to mutual understanding and learning (Habermas, 1990). Rather than publishing thousands of data points or lines of code, we need a transparency that helps justify why certain policy decisions are acceptable and fair. The key seems to be an open dialogue between decision-makers and the subjects (and wider publics) affected by these decisions, regarding which knowledge claims will be deemed valid and under which circumstance. Practically, this could mean justification based on the predictive power of a model, the underlying modelling assumptions, good model fit, or the scientific rigour involved in developing the model (Binns, 2018).

As this tentative discussion illustrates, there is an urgent need for wider critical reflection on the regulatory implications of a Big Data State of Mind. A task that has barely begun despite the speed of change in the data landscape. In this chapter I have demonstrated that the shift to a Big Data State of Mind alters the behaviour of regulators, and their relationships with regulated entities, whilst (re)making our perceptions of the world along with it. We must consider further how performance indicators participate in shaping the world with us as we use them.

Are stress tests breaking the bank?

This chapter addresses this question as whether the stress test can help us understand which banks are 'broken', and why. Just as credit ratings before the crisis wrongly told us that certain assets were safe (Rafailov, 2011), it is not automatically guaranteed that the stress test is right when it tells us that certain banks are healthy.

In the stress test, supervisors vet banks' results through a (heavily data-driven) top-down Quality Assurance process. When there is a discrepancy between banks' results and supervisor's results, banks' results are often overridden. However, respondents complained that it is unclear how these results are obtained, and what they mean for banks' individual risk. Because supervisors do not engage in any discussion with banks to justify their results, this undermines the credibility of the exercise.

Overall, I would conclude that it is difficult to tell why certain banks 'break' in the stress test and others do well. This is not necessarily problematic, as large data sets are designed to predict, rather than understand, outcomes. However, for the stress test, the validity of these models, and thus the validity of the results is called in question. This is the true problem of the stress test. The exercise could certainly be enriched by a better dialogue between regulators, banks, and wider publics to justify the validity of the knowledge claims made.

CONCLUSION

Performance Indicators have intruded into the fibres of government (Christopher Hood, 2007). What makes them so popular is their ability to reduce complexity and allow information to be processed easily. They can be used to set targets, foster competition, allocate scarce resources, and hold underperformers accountable. We rely on indicators to tell us how our societies are doing. Indicators enable us to monitor performance at a glance, increasing transparency and informing choice (Boswell 2018). Given the proliferation of performance indicators in society, it is imperative to have a good understanding of where this performance information comes from, and what it does.

Although performance indicators have received substantial academic attention, the literature so far has been overly focused on the use, users, and non-use of performance information (Van Dooren et al., 2015). Consequently, we know little about how performance indicators are made. When studied, the design of indicators is mainly dealt with as a technical matter, discussing analysis techniques and other methodological issues (see for instance McGlynn & Asch, 1998; van Hoek, 1998). The socio-political process of knowledge production remains black-boxed and largely unknown to us (important exceptions include Bartl et al., 2019; Berten, 2019; Cook, 2017; Davis et al., 2012b; Espeland, 2015; Gorur, 2015).

To further the understanding of the regulatory implications of processes of knowledge production through indicators, I conducted a critical and in-depth case study of the EU-wide banking stress test, following an interpretive methodology. The stress test is an indicator that provides information on the health of systemically important banks in the EU by projecting how they would perform in a hypothetical crisis scenario. The results make it easy to compare banks vis-à-vis each other, and rank them according to their performance. The results of the stress test feed into banks' Supervisory Review and Evaluation Process (SREP), which amongst others determines their capital decision. I interviewed a wide array of people involved in the design and calculation of the stress test. Respondents included supervisors and experts at the European Central Bank (ECB) and European Banking Authority (EBA), National Competent Authorities at the National Bank of Belgium (NBB), risk directors and members of the stress testing teams in Belgian banks, as well as consultants involved in the process. This allowed me to gain a comprehensive understanding of how the stress test was made, how it changed over time, and how this affected regulators, regulatees, and their relationships.

Considering the different findings across the empirical chapters, it is possible to formulate an answer to the research question '**What are the regulatory implications of knowledge production through indicators?**'. Overall, I argue that the process of producing indicators is regulatory in itself, i.e. it steers behaviour. I look more closely at how indicators measure, manage and make, arguing that indicators steer behaviour through largely unrecognized processes of commensuration, governmentality, and epistemic

shifts. Based on the results of this dissertation, it is possible to weigh in on some long-standing debates, as well as integrate and add to current research. In the next section, the key findings are discussed. Then, I reflect on the case of the stress test; if and how it is **'breaking the bank'**, i.e. is the stress test a good regulatory tool? Finally, I contemplate the wider implications of this research, and suggest several practical recommendations and avenues for further research.

KEY FINDINGS

As mentioned in my methodological chapter, my findings are not generalisable in a law-like, statistical way. It is not because I find that the stress test treats banks as incommensurables, that other indicators necessarily do the same. Rather, I provided a contextualised understanding of why this is the case, and how it is done. It is the causal patterns underlying the *why's* and the *how's* in the calculation process that are generalisable to other settings. By explaining how certain events were triggered in this context, it becomes clear how and why they might be triggered in others. In what follows I elaborate on the three key lessons that can be learned from this research.

HOW INDICATORS MEASURE

Lesson 1: measurement 'bias' can be intentional; it is not always the product of capture, gaming, or industry lobbying. Instead, regulators can choose to consciously introduce bias for strategic reasons, a.o. to encourage learning and innovation.

This dissertation addresses the notion of measurability. I follow Michel Callon and Fabian Muniesa's (2005) idea that in order to be measured, entities must be made measurable first. In the case of indicators, entities do not only need to be made measurable, but also made comparable. This is done through the process of commensuration, i.e. taking diverse qualitative entities and homogenizing them quantitively on a common metric (Espeland & Stevens, 1998). Regulators are usually in favour of commensuration, while regulated sectors typically fight these commensuration efforts in defence of their unique qualities (Gorur, 2016; Peeters et al., 2014; Porter, 1995; Scott, 1998). Efforts to withstand commensurability (such as regulatory capture, lobbying, and gaming), distort the validity and comparability of the results, introducing (what is often seen as) unwanted measurement bias.

One important finding is that bias is not always a sign of concession to industry interests. Instead, leaving room for incommensurability can serve regulators' interests as well. For instance, by not adhering to a predefined shared definition or model of how risk should be measured, regulators share the accountability for this modelling and measurement with banks. More importantly, incommensurability leaves room for what Barry (2012) calls 'knowledge controversies', a continuous debate and reconsideration regarding how (in this case) risk should be understood and calculated. This

carries important learning opportunities for both regulators and regulatees, and fosters innovation. In the case of the stress test, banks are challenged to not just copy-paste models provided by regulators, but to invest in in-house expertise to develop their own models. To be sure, it must be emphasised that the findings of this study do not wholly refute the role of capture, lobbying or gaming in distorting indicator outcomes. Rather, it highlights that in some circumstances, bias can also be a conscious choice, and that these mechanisms are not as pervasive as often posited.

Moreover, I challenge the idea that some things are by nature easier to measure and more commensurable than others. Processes of commensuration are often overlooked and underestimated in policy fields like the financial sector, where quantification is taken for granted due to wide-spread standardizing mechanisms such as double-entry bookkeeping and International Financial Reporting Standards (IFRS). Commensuration can seem natural when measurement systems have become widely shared conventions, or where they are black boxed. I find it quintessential to especially challenge measurement practices where they are taken for granted.

How indicators manage

Lesson 2: Performance management by indicators does not only occur through management-by-results. Rather, the process of calculating the indicator can be regulatory in itself, i.e. it can cause important long-term changes in attitudes, habits and beliefs.

This dissertation makes important contributions to the current debates on performance management. Performance management is often seen as the

process of putting performance measures to practice, i.e. using the performance information to regulate behaviour and improve performance outcomes. Whether performance indicators deliver on this promise, is increasingly up for debate. Although many success stories exist on how performance measurement positively affects performance outcomes (Boyne & Chen, 2006; Nielsen, 2014; Walker et al., 2011), in the past years, performance indicators have had to endure severe criticism. They are said to lack accuracy, encourage tunnel vision, create perverse incentives, and ultimately fail to improve performance. Even sympathetic analysts, like Hood & Peters (2004) or Dunleavy et al. (2005) acknowledge the adverse effect of performance indicators (Pires, 2011). Which raises the question why performance indicators improve performance outcomes in some cases and fail to do so in others.

I argue that performance measures do not only manage by results, but already manage behaviour through the design of the measurement process. I find that the design of the indicator can have an impact on regulatees' attitudes and encourage processes of self-regulation. I link this to Foucault's notion of governmentality (Foucault et al., 1991; Lemke, 2011; Rose, O'Malley, & Valverde, 2006). In my case, banks are not likely to change their behaviour based on the results of the indicator; they even find this information to be invalid. Instead, through the process of having to produce this knowledge, they configure new habits and patterns of behaviour, encouraging processes of self-regulation. I draw from this that indicators can be latently constitutive of the performances they intend to measure. For instance, banks updated IT-systems, increased communication across

departments, and improved internal processes. Although they initially only made these changes to be able to calculate the performance indicator, these changes improved their internal risk management systems. The process of calculating the indicator was regulatory in itself, i.e. it caused important long-term changes in attitudes, habits and beliefs.

Performance indicators do not only steer the behaviour of regulated entities. It has been widely demonstrated that they can be used to manage third party expectations and keep up appearances to a wider public (Boswell, 2008). My work also demonstrates how such processes can work performatively, meaning that they engender that which they describe (Callon, 2010). My research highlights that these processes do not work as self-fulfilling-prophecies, rather they require work in order to act performatively. A statement of good performance needs to be backed by credible institutions, a credible measure, and over time by improved performance. It is only when the stress test got taken over by the ECB, and the ECB made it a far more thorough exercise, with a fifty instead of five page rulebook, that market's faith began to be restored (Anderson, 2016). And it was only after banks stopped passing the exercise but faltering in real life shortly after[52], that the exercise was taken more seriously.

[52] Take for instance the case of the Irish banks in need of a bailout in 2010, shortly after passing the CEBS 2009 stress test (Schneibel & Braun, 2010).

How indicators make

Lesson 3: Indicators can (re)make how we understand the world, and thus how we govern it. We increasingly rely on data-driven decision-making, which warrants new frameworks of accountability.

This dissertation reflects on how indicators make. Indicators are typically seen as measures that describe the world we live in. However, a case can be made that indicators do more than simply summing up pre-existing characteristics of the world. They can actively (re)make the way we see the world, and what we accept as truth.

I argue there is a shift towards indicators based on large data sets, which is accompanied by an epistemic shift to what I call a 'Big Data State of Mind'; the epistemological notion that one can or should rely on large data sets rather than theory to observe and understand reality. This is not necessarily problematic, algorithms based on large data sets might do a much better job at describing the world for us, because humans are biased and theories can be flawed (Anderson 2008). On the other hand, algorithms may also be trained on biased data and continue to discriminate (O'Neil, 2016). As such, it is important to be able to justify data-driven decision-making, and hold policy makers accountable for these decisions.

Traditional notions of accountability and transparency are not always compatible with a Big Data State of Mind. Sharing lines of code and intricate data matrixes is not always possible, for instance to avoid gaming (Bambauer, 2017), nor is it necessarily useful, because it still would not shed any light on why the algorithm is making certain transformations (Diakopoulos, 2016;

Pasquale, 2015). If the eventual aim is to ensure fair decision-making, this is better resolved through a different kind of transparency. Drawing on insights from chapter four, it would prove more helpful to view transparency from a framework of communicative rationality (Habermas 1990); as a tool to increase understanding through communication. A dialogue should be opened between supervisors, supervised entities (and possibly also wider publics), to discuss which knowledge claims will be accepted as valid, and under which conditions. Decisions based on large data sets might be justified by the scientific rigour with which algorithms are developed, the quality standards they adhere to, and their predictive power. Ultimately, we might have to learn to accept policy decisions we do not necessarily understand.

BREAKING THE BANK?

In 2007 and 2008 the United States and Europe were plunged into the biggest financial crash and economic recession since the nineteen thirties. The trigger for previous major crises had usually been some sort of shock, like the oil price spike or war. This time, collapse was linked to the malfunctioning of the financial system itself. Just before the crash, all the standard financial indicators declared the financial system safe. Banks were awarded top notch credit ratings and deemed sufficiently capitalized. National regulators seemed to be doing their job well (Farlow, 2015). And then, as if from nowhere, the banking sector broke and millions of citizens were in danger of losing their jobs, savings, and even their homes.

In the EU and the US supervisors started working on new financial stability strategies, one of which was the deployment of stress tests. Supervisors would take a thorough look at the portfolios of systemically important banks and project how they would fare under a severely adverse scenario. As Timothy Geithner (2015, p. 17) writes: "The stress test would provide a form of triage, separating the fundamentally healthy from the terminally ill". Unfortunately, especially in the EU, the stress tests have not been exempt from criticism, mainly regarding the severity and credibility of the results (Dowd, 2015).

ARE THE STRESS TESTS TOUGH ENOUGH ON BANKS?

This warrants a discussion whether the stress tests are stressful enough. Are they really 'breaking' the bank, or are they just communication exercises? Much can be said about the shortcomings of the stress test. For instance, the crisis scenario that the EU banks were subjected to was deemed far less severe than that of the Federal Reserve or the Bank of England, leading to misleadingly positive results for the European Banking Sector (Cecchetti & Schoenholtz, 2016). This was put into context by a respondent working at the ECB. He used the sentence 'you're stuck with what you can afford' to explain how the reactions the US could afford, and the reactions Europe could afford were very different. In the US, failing banks who could not attract enough private capital to bridge the gap, were given a capital injection from the government (Geithner, 2015). In Europe, there were no such backstops. Respondents explained that if European banks would be seen to lack similar amounts of capital as their US counterparts, markets would be in a frenzy and

even less likely to inject capital into the financial system, only aggravating the crisis. A mechanism of performativity, that I described earlier. Respondents emphasised that it was not capture or lobbying, that caused the scenario to be less stressful for banks, it was part of regulators' strategy to contain the crisis.

European supervisors have also been criticised regarding the bottom-up approach they took (Goldstein, 2015). Banks were allowed to use their own models to calculate the indicator, rather than all banks taking a uniform approach imposed by the regulator (top-down). This was said to leave the exercise susceptible to gaming efforts[53], and again not very tough. However, I find that this approach carries important benefits. Most importantly, it encourages banks to develop in-house risk management expertise to create models tailored to their assets. It also stimulates more discussion and deliberation between regulators and regulatees regarding how risk should be measured and understood. In what are called 'narratives', banks explain to regulators why they model the impact of the macro-economic scenario on their portfolio in a specific way. This gives regulators a fine-grained understanding of banks and forms a good starting point for discussion (at least theoretically[54]). As such, the European bottom-up approach allows for what Barry (2012) calls knowledge controversies; a continuous reconsideration of how to understand and measure risk.

[53] It is also important to note here that these internal models have been previously vetted and approved by supervisors.

[54] Respondents complained that actual discussions with supervisors were rare. If they had remarks or questions, they passed it on to their JST's, who then passed it on to the DG4 modelling experts, who then passed on their response through the JST's. Although the JSTs are an important intermediate to buffer gaming efforts, this system was not optimal to stimulate learning and innovation.

The answer to the question whether stress tests are stressful enough in the EU is a complicated one. It is fair to say that the macro-economic scenario was less stressful than its US counterpart. However, it is important to qualify that this was not because European regulators were captured by industry interests as critics are eager to claim. Moreover, the bottom-up approach did leave more room for gaming, but also for learning opportunities. This means the stress test might be tough enough; but is it worth all the effort?

DO THE BENEFITS OF THE STRESS TEST OUTWEIGH THE COSTS?

A second point of criticism deals with the high cost of the stress test in terms of public and private resources (Thun, 2013). This begs the questions if the stress tests are metaphorically 'breaking the bank'? Are they too costly for what they deliver, or do the benefits outweigh the many resources invested in them? It is clear that the development of the stress test in Europe has been a learning-by-doing exercise. For each exercise supervisors still make changes to the methodology and data requirements. This makes it costly for banks who find it hard to automate processes and find some routine in the exercise. Respondents in banks found it especially frustrating that they were required to invest their resources in providing granular information, for which they saw no real purpose.

However, in order to access this granular data from subsidiary branches, banks were encouraged to optimise their internal risk management systems. For instance, they reconciliated data across diverse systems, built IT-tools to automate large parts of the stress testing exercise, and created more

formalized communication structures across risk departments. These changes were only possible because CEO's were more inclined to invest in risk departments with their reputation at stake in the stress test. The changes did not only help banks complete the stress testing exercise more efficiently, but they also continued to prove their use in banks' day-to-day risk management. The stress test thus led banks to improve their self-regulation and created an important incentive for CEO's and boards of directors to invest more in their banks' risk departments.

When asked the question if the benefits outweigh the costs, I would argue that in order to weigh the costs and benefits, one needs to consider the wide array of benefits that span beyond the scope of the stress test. Moreover, as the regulators further optimise the exercise, it is likely that the costs will be reduced. To be sure, beyond further improvements in the methodology of the stress test, important reconsiderations are left to be made in terms of the general philosophy of the exercise, as will be elaborated in the next point.

DOES THE STRESS TEST HELP US UNDERSTAND WHICH BANKS ARE 'BROKEN' AND WHY?

A final reflection can be made on how the stress test helps us understand which banks are broken and why. How does the stress test decide if a bank is healthy or not? Financial institutions worldwide are facing an increased level of regulatory scrutiny in the aftermath of the crisis. Banks are now asked to provide granular data for their risk assessment in general, and more

specifically also for stress testing exercises. Van Steen (2015) calls these requirements 'regulatory big data'. The increased quantity of data presents an opportunity for regulators to enhance their understanding of the institutions they regulate. Where regulators used to ask for specific figures that spanned an excel file or two every quarter, they have become far more scrupulous since the crisis. As interviews at the National Bank and the ECB showed, supervisors are unsure where the next crisis is going to come from. As such, data requirements increase with every stress test because no one wants to overlook what might be important information. Unsure as to where to look, supervisors try to look everywhere.

Although the exercise has a bottom-up nature (banks calculate their own results), supervisors also use a (data-driven) top-down Quality Assurance (QA) process to vet these results, often overriding banks' outcomes. This makes it especially difficult to gauge where the final results come from and what they mean. Moreover, supervisors and risk managers in banks raised concerns that despite the large amounts of data, regulators were still not able to pin point which banks were actually doing well in real life. For instance, banks that did not de-risk their portfolio, still scored well in the exercise.

When asked the question if the stress test contributes to a better understanding of why certain banks 'break', I would say no. This is not necessarily problematic, as large data sets are designed to predict, rather than understand, outcomes. However, for the stress test, the predictive power and the validity of the models used by supervisors is also called in question. Especially as supervisors refrain from discussing the models they use in the QA process. The lack of justification of the results is the real

problem of the stress test. This could be dealt with through a more extensive dialogue between supervisors, supervised entities, and wider publics.

In conclusion, I would argue that although the stress test scenario might not have been extremely challenging for banks, allowing for rather optimistic results, it does create learning opportunities, and contribute to the professionalization of risk management in banks. Meaningful steps have been made to "assess the resilience of financial institutions" (EBA, 2019a), yet, important work is left to justify the results of the exercise.

A FUTURE FOR PERFORMANCE MANAGEMENT: WIDER IMPLICATIONS, RECOMMENDATIONS, AND AVENUES FOR FURTHER RESEARCH

With the quintessential so what question in mind; it is important to specify the wider implications of a research project. Moreover, it is useful to give some indications or direction of what can or should be done to improve the current approach to performance management, and where further research can be helpful. Ultimately, I argue we need to change our expectations about how performance management succeeds.

COMMUNICATIVE RATIONALITY AND LEGITIMACY

First of all, there is no 'one best way' to design a performance indicator. Knowledge production is more than a technical process, it is a social one. Designing an indicator should be about more than finding a supposedly objective way to measure a phenomenon. It is problematic to view the measurement process as purely technical. This approach promotes an instrumental conception of knowledge, reminiscent of Habermas' (1984) notion of instrumental rationality, which is strategic and results oriented. Instrumental knowledge aims to manipulate the world, by imposing 'rational' knowledge, in order to control it (Marcuse, 1941). In this interpretation, performance information is conceived of as a fixed outcome of a so-called rational process, that can then be used to affect behaviour through rules on permissible deviations from a standard.

I argue it is more helpful to approach the design of indicators according to Habermas' logic of communicative rationality (1984). This approach focuses more on increasing understanding through open communication. Here, knowledge production is reconceived as an ongoing exchange among critical equals, rather than a rational process imposed by dominant actors. Designing performance indicators in this non-instrumentalised way establishes discursive conditions that offer a more critical and understanding-oriented space for knowledge production. Practically this means that when designing performance indicators, processes of commensuration should not be seen as panacea. Commensuration can stifle learning opportunities and hamper knowledge building processes. Instead I argue for a bottom-up approach, and a collaborative design to

indicators, where there is room to negotiate incommensurability. Such an approach has the potential to engage regulated entities and regulators in a creative and democratic construction of indicators. Stone (2012, p. 284) calls such indicators, developed in a participative way, 'active indicators'. This type of indicator suggests an iterative approach, where dialogue is crucial. Not only technical dialogue regarding data availability and statistical methods, though this is important too, but a substantive and normative dialogue regarding prioritisation and conceptualisation. Especially when it comes to measurement in fields of high uncertainty and contestation, dialogue among stakeholders can be crucial to gain an encompassing understanding of the issue at hand.

This can also help to improve the legitimacy of (global) indicators. Stone (2012) identifies the weak sources of legitimacy for the standards implicit in indicators as a persistent problem in the construction of indicators. Questions of legitimacy were raised in the stress test as well. For instance, in the US the Federal Reserve took a more top-down approach to the stress test. Although this is commonly seen as more successful in tackling potential gaming efforts, it is also alleged to be more opaque, and lead to uncertainty among banks who are seeking for more transparency in the methodology imposed by regulators (Enria, 2018). This ties in with my findings regarding the benefits of the bottom-up calculations in the EU, where joint knowledge production is seen to support the legitimacy of the indicator. Still, there is more work to be done to study how top-down and bottom-up calculation processes affect legitimacy. For instance, top-down calculations may seem opaquer, but they might also succeed better in establishing a level playing

field, which is seen to improve legitimacy (Levitsky & Lucan, 2009). Additionally, bottom-up calculations require intensive cooperation, which is more susceptible to power plays, hampering legitimacy (van Dijk, 2008).

GOVERNMENTALITY, VALIDITY, AND RELIABILITY

Secondly, performance indicators are not all about the numbers. Providing valid and reliable information is often seen as the main goal of performance indicators (Hood, 2007). The general assumption of most performance management research is that in order to steer performance outcomes, we need good performance information (Moynihan, 2008; Pollitt, 2018; Van Dooren et al., 2015). We need to know the inputs, outputs, throughputs, and outcomes, to manage government and public services. An increasing number of critical voices claim that in many cases performance indicators fail as management tools, because they fail to provide such valid and reliable information, as they are gamed or manipulated (Davis et al., 2012a; Dunleavy et al., 2005; C. Hood & Peters, 2004; van Thiel & Leeuw, 2002).

I argue that performance indicators can still succeed as management tools, even if they fail to provide valid or reliable information[55]. The measurement process can affect performance outcomes, regardless of the performance results and their reliability or validity. Rather than learning from results, organisations can learn from producing the information. In the case of the stress test, despite the fact that banks found the results of the exercise

[55] To be sure, the costs still need to outweigh the benefits; I simply argue that latent benefits should not be overlooked.

invalid, calculating the exercise still caused them to professionalise their own surveillance and risk management.

I tie this to Foucault's notion of governmentality (Foucault, 2011; Lemke, 2011). Rather than directly regulating behaviour (requiring banks who underperform on the stress test to recapitalize), the stress test steers how banks self-regulate (banks decide themselves to professionalise their own surveillance and risk management in order to be able to calculate the stress test), ultimately improving performance outcomes. Governmentality mechanisms can play an important part in organisational reform, and have so far been largely overlooked in literature (Brunsson & Olsen, 2018).

More research is necessary to determine the specific conditions in which this organisational reform occurs. A key reason that banks decided to improve their internal risk-management systems, was the high visibility and consequence of the indicator. Banks were highly motivated to complete the exercise in a timely fashion. For less visible, or less consequential indicators, organisations may not be as likely to reform. Moreover, how the calculation process should be designed to instigate this self-regulation, is still unclear. In this case extensive data-requirements were key to initiate change, in other contexts it may be more appropriate to introduce other design parameters to trigger certain behavioural responses. Extensive data requirements may sometimes create unnecessary red tape, and take away time from other tasks.

A final important remark is that steering through governmentality, rather than results, can also help to avoid decoupling; organisations claiming to fulfil requirements while internally not changing anything (Meyer & Rowan, 1977). In steering by results, organisations may try to circumvent actual

change by window dressing or gaming results rather than actually practicing organisational reform. However, by developing an exercise that required banks to have an effective management system in place in order to complete it, banks were left no choice but to de facto improve their internal risk-management. Overall, using performance indicators to steer behaviour through mechanisms of governmentality, rather than solely through management-by-results, holds many benefits and seems a promising avenue for practitioners to explore.

DATA-DRIVEN INDICATORS AND ACCOUNTABILITY

Finally, we need to rethink how indicators can succeed as tools of governance in an era of 'Big Data', especially in terms of accountability. Accountability and transparency are often identified as founding principles of public administration, closely related to well-functioning democracy (Dubnick & Frederickson, 2011; Keane, 2009; Mulgan, 2003). Accountability is even described as the über-concept of the twenty-first century (Flinders, 2014). Though accountability is seen to mean many things, generally it refers to an obligation to explain or justify one's actions (Bovens, 2010). The conventional wisdom amongst policy scholars is that transparency generates accountability (Fox, 2007). Policy makers are held to account through detailed disclosure of their actions. When governments come up with an indicator score, it needs to be clear where this score comes from, how it was calculated. Throughout history, transparency, as an ideal, has offered a way to see inside the truth of government (Ananny & Crawford, 2018).

Yet, full transparency is not always meaningful and it can often prove problematic. For instance, the open publication of hundreds of thousands of data points can be seen as a good step in terms of accountability. However, as a respondent noted, too much information kills information. Without a straightforward way to make sense of this data, it is rather meaningless. Transparency should be about more than simple data-disclosure, as also advocated by Winkler (2000). More information does not always improve the clarity of information, and can be of little use to hold policy makers accountable. Moreover, full disclosure is not always feasible or desirable because it enables gaming efforts which could be destructive to the regulatory system, as argued by respondents at the ECB and confirmed in literature (Bambauer, 2017).

To be sure, the (intentional) obscuring of decision-making processes is equally harmful to a democratic system, as it may lead to abuse of office and unfair assessments (Christopher Hood, 1991). This becomes especially salient when decision-making is increasingly data-driven, and algorithms come into play. Vedder and Naudts (2017) point out that it often remains unclear why algorithms make certain transformations, and no amount of transparency regarding lines of code or data matrixes will change that. As such, traditional accountability mechanisms cannot be appropriately applied to algorithms operating on Big Data.

Ananny and Crawford (2018) importantly add to this that seeing does not always mean knowing. This ties in with the constructivist assumptions of this dissertation. Truth is understood as relational, created in interaction, rather than something that 'is'. The only way to see truth then is through its

'becoming' in interaction, rather than through what it supposedly 'is'. As such, transparency is not a state in which things 'are' clear, but a state in which things are 'made' to be clear (in a given way), transparency is performative rather than descriptive.

Rather than looking 'inside' decision-making systems, we should hold these systems accountable by looking 'across' them: By seeing them as sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans. Transparency is not simply about revealing information, but deploying devices, actor-networks, that manage 'visibility'. An algorithm is not just code, but an actor-network where code, people, and platforms intersect and continuously interact. If the truth is not a positivist discovery, but a relational achievement, the target of transparency should be relational. Instead of asking what something 'is', the question should be how it is made to be so.

Arguably, in assessing how the stress test is made, this dissertation is an example of such relational transparency. It has painted a picture of the vast regulatory implications of knowledge production through indicators.

REFERENCES

Abolafia, M. Y. (1998). Markets as Cultures: An Ethnographic Approach. *The Sociological Review*, *46*(1), 69–85. https://doi.org/10.1111/j.1467-954X.1998.tb03470.x

Ackerman, F., & Heinzerling, L. (2004). *Priceless: On Knowing the Price of Everything and the Value of Nothing*. New York: The New Press. https://doi.org/10.1136/bmj.330.7499.1091

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete | WIRED. Retrieved from https://www.wired.com/2008/06/pb-theory/

Anderson, R. W. (2016). *Stress Testing and Macroprudential Regulation A Transatlantic Assessment*. London: CEPR Press.

Bambauer, D. E. (2017). Uncrunched: algorithms, decision making, and privacy. In *second annual digital information policy scholars conference*. Arlington, VA: George Mason University Antinin Scalia Law School.

Barry, A. (2012). Political situations: knowledge controversies in transnational governance. *Critical Policy Studies*, *6*(3), 324–336.

https://doi.org/10.1080/19460171.2012.699234

Bartl, W., Papilloud, C., & Terracher-Lipinski, A. (2019). Governing by Numbers - Key Indicators and the Politics of Expectations. An Introduction. *Historical Social Research / Historische Sozialforschung*. GESIS - Leibniz Institute for the Social Sciences. https://doi.org/10.2307/26604897

Barzelay, M., & Armajani, B. J. (1992). *Breaking through bureaucracy : a new vision for managing in government*. Cambridge, UK: University of California Press.

Baudrillard, J. (1994). *Simulacra and simulation*. Michigan: University of Michigan Press.

Bauman, Z. (1989). *Liquid modernity*. Polity Press.

Baxter, L. G. (2011). Capture in Financial Regulation: Can We Channel It toward the Common Good. *Cornell Journal of Law and Public Policy*, *211*(January), 175–200.

Belpaire, A. (1847). *Traité des dépenses d'exploitation aux chemins de fer*. Antwerp: J.E. Buschmann.

Bennett, J. (2011). *Modernity and its Critics*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199604456.013.0006

Berten, J. (2019). Failed Indicatorisation: Defining, Comparing and Quantifying Social Policy in the ILO's International Survey of Social Services of the Interwar Period. *Historical Social Research / Historische*

Sozialforschung. GESIS - Leibniz Institute for the Social Sciences. https://doi.org/10.2307/26604903

Berten, J., & Leisering, L. (2017). Social policy by numbers. How international organisations construct global policy proposals. *International Journal of Social Welfare*, *26*(2), 151–167. https://doi.org/10.1111/ijsw.12246

Bevan, G., & Hood, C. (2006). What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration*, *84*(3), 517–538. https://doi.org/10.1111/j.1467-9299.2006.00600.x

Bevir, M. (2010). Rethinking governmentality: Towards genealogies of governance. *European Journal of Social Theory*, *13*(4), 423–441. https://doi.org/10.1177/1368431010382758

Bijker, W. E., & Law, J. (1992). *Shaping technology/building society : studies in sociotechnical change*. Massachusetts: MIT Press.

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Bloor, D. (1981). Sociology of (scientific) knowledge. In J. Browne, W. Bynum, & R. Porter (Eds.), *Dictionary of the history of science* (pp. 391–393). Chicago, IL: University of Chicago Press.

Bohte, J., & Meier, K. J. (2000). Goal Displacement: Assessing the Motivation for Organizational Cheating. *Public Administration Review*,

60(2), 173–182. https://doi.org/10.1111/0033-3352.00075

Boswell, C. (2008). The political functions of expert knowledge: knowledge and legitimation in European Union immigration policy. *Journal of European Public Policy*, *15*(4), 471–488. https://doi.org/10.1080/13501760801996634

Boswell, C. (2015). The double life of targets in public policy: disciplining and signalling in UK asylum policy. *Public Administration*, *93*(2), 490–505. https://doi.org/10.1111/padm.12134

Boswell, C. (2018). *Manufacturing Political Trust*. Cambridge University Press. https://doi.org/10.1017/9781108367554

Boswell, J. (2018). What makes evidence-based policy making such a useful myth? The case of NICE guidance on bariatric surgery in the United Kingdom. *Governance*, *31*(2), 199–214. https://doi.org/10.1111/gove.12285

Bouckaert, G., & Balk, W. (1991). Public Productivity Measurement: Diseases and Cures. *Public Productivity & Management Review*, *15*(2), 229. https://doi.org/10.2307/3380763

Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics*, *33*(5), 946–967. https://doi.org/10.1080/01402382.2010.486119

Bowen, G. A. (2006). Grounded Theory and Sensitizing Concepts. *International Journal of Qualitative Methods*, *5*(3).

Boyd, D., & Crawford, K. (2012). Critical Questions For Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Boyne, G. A., & Chen, A. A. (2006). Performance Targets and Public Service Improvement. *Journal of Public Administration Research and Theory*, *17*(3), 455–477. https://doi.org/10.1093/jopart/mul007

Braithwaite, J. (2014). Restorative justice and responsive regulation: the question of evidence. *RegNet Research Paper*, *51*, 27. https://doi.org/http://dx.doi.org/10.2139/ssrn.2514127

Bruno, I., Jacquot, S., & Mandin, L. (2006). Europeanization through its instrumentation: benchmarking, mainstreaming and the open method of co-ordination … toolbox or Pandora's box? *Journal of European Public Policy*, *13*(4), 519–536. https://doi.org/10.1080/13501760600693895

Brunsson, N., & Olsen, J. P. (2018). *The Reforming Organization: Making sense of administrative change*. London: Routledge. https://doi.org/10.4324/9781351252188

Brunsson, N., & Sahlin-Andersson, K. (2000). Constructing Organizations: The Example of Public Sector Reform. *Organization Studies*, *21*(4), 721–746. https://doi.org/10.1177/0170840600214003

Bryer, R. A. (2000). The history of accounting and the transition to capitalism in England. Part one: theory. *Accounting, Organizations and Society*, *25*(2), 131–162. https://doi.org/10.1016/S0361-3682(99)00032-X

Bunea, A. (2013). Issues, preferences and ties: determinants of interest groups' preference attainment in the EU environmental policy. *Journal of European Public Policy*, *20*(4), 552–570. https://doi.org/10.1080/13501763.2012.726467

Burr, V. (1995). *An Introduction to Social Constructionism*. Abingdon-on-Thames: Routledge. https://doi.org/10.4324/9780203133026

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 176–198. https://doi.org/10.1177/2053951715622512

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, *32*(1), 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Callon, M. (1991). Techno-economic networks and irreversibility. In J. Law (Ed.), *Sociology of Monsters: Essays on Power, Technology and Domination* (pp. 132–161). London: Routledge.

Callon, M. (2010). Peformativity, Misfires, and Politics. *Journal of Cultural Economy*, *3*(2), 163–169. https://doi.org/10.1080/17530350.2010.494119

Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an uncertain world: an essay on technical democracy*. MIT Press.

Callon, M., & Muniesa, F. (2005). Peripheral Vision: Economic Markets as Calculative Collective Devices. *Organization Studies*, *26*(8), 1229–1250.

https://doi.org/10.1177/0170840605056393

Cecchetti, S. G., & Schoenholtz, K. L. (2016). Are European Stress Tests Stressful Enough? *Money & Banking*. Retrieved from https://www.moneyandbanking.com/commentary/2016/8/8/are-european-stress-tests-stressful-enough

Chan, J., & Moses, L. B. (2016). Is Big Data challenging criminology? *Theoretical Criminology*, *20*(1), 21–39. https://doi.org/10.1177/1362480615586614

Christensen, T., & Fan, Y. (2016). Post-New Public Management: a new administrative paradigm for China?: *International Review of Administrative Sciences*, *84*(2), 389–404. https://doi.org/10.1177/0020852316633513

Christensen, T., & Lægreid, P. (2006). *Autonomy and regulation : coping with agencies in the modern state*. London: Edward Elgar. Retrieved from https://books.google.be/books/about/Autonomy_and_Regulation.html?id=ZYP4ngEACAAJ&redir_esc=y

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data : complementary research strategies*. Thousand Oaks, CA: Sage.

Cook, E. (2017). *The pricing of progress : economic indicators and the capitalization of American life*.

Czarniawska, B. (2014). Following Objects and Quasi-objects. In *Social Science Research: From Field to Desk* . London: Sage. Retrieved from

https://www.amazon.com/Social-Science-Research-Field-Desk-ebook/dp/B00K21JKN0/ref=cm_cr_arp_d_pdt_img_top?ie=UTF8

Davis, K. E., Kingsbury, B., & Merry, S. E. (2012a). *Governance by indicators : global power through quantification and rankings*. Oxford: Oxford University Press [in association with] Institute for International Law and Justice, New York University School of Law.

Davis, K. E., Kingsbury, B., & Merry, S. E. (2012b). Indicators as a Technology of Global Governance. *Law & Society Review*, *46*(1), 71–104.

De Bièvre, D., & Bursens, P. (2017). Patterns of Covert Integration in EU Governance. A Response to Adrienne Héritier. In P. Bursens, C. De Landtsheer, L. Braekmans, & B. Segaert (Eds.), *Complex Political Decision-making: Leadership, Legitimacy, and Communication*. Farnham: Ashgate.

de Bruijn, H. (2001). *Prestatiemeting in de publieke sector: tussen professie en verantwoording*. Utrecht: Lemma.

De Bruyckere, V., Gerhardt, M., Schepens, G., & Vander Vennet, R. (2013). Bank/sovereign risk spillovers in the European debt crisis. *Journal of Banking & Finance*, *37*(12), 4793–4809. https://doi.org/10.1016/J.JBANKFIN.2013.08.012

de Maria, B. (2008). Neo-colonialism through measurement: a critique of the corruption perception index. *Critical Perspectives on International Business*, *4*(2/3), 184–202.

De Ruiter, R. (2008). Developing Multilateral Surveillance Tools in the EU. *West European Politics*, *31*(5), 869–914.

de Vries, G. (2016). *Bruno Latour*. Cambridge: Polity.

de Vries, M. S. (2010). Performance measurement and the search for best practices. *International Review of Administrative Sciences*, *76*(2), 313–330.

Denhardt, R. B. (1981). Toward a Critical Theory of Public Organization. *Public Administration Review*, *41*(6), 628. https://doi.org/10.2307/975738

Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*, *49*(7), 1043–1064. https://doi.org/10.1177/0095399714555751

Desrosières, A. (1993). *The politics of large numbers : a history of statistical reasoning*. Cambridge, MA: Harvard University Press.

Desrosières, A. (2015). Retroaction: how indicators feed back onto quantified actors. In R. Rottenburg, S. E. Merry, S.-J. Park, & J. Mugler (Eds.), *The World of Indicators* (pp. 329–353). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781316091265.013

Diakopoulos, N. (2016). A view from computational journalism. *Communications of the ACM*, *59*(2). https://doi.org/10.1145/2844110

Dickens, C. (1854). *Hard times*. London: Bradbury & Evans.

Doig, A., Mcivor, S., & Theobald, R. (2006). Numbers, nuances and moving targets: converging the use of corruption indicators or descriptors in assessing state development. *International Review of Administrative Sciences*, *72*(2), 239–252. https://doi.org/10.1177/0020852306064612

Dowd, K. (2015). Central Bank stress tests : mad, bad, and dangerous. *Cato Journal*, *35*(3), 507–524.

Drisko, J., & Maschi, T. (2015). *Content Analysis*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190215491.001.0001

Drucker, P. F. (Peter F. (1968). *The age of discontinuity : guidelines to our changing society*. Oxford: Butterworth-Heinemann.

Dubnick, M., & Frederickson, H. G. (2011). *Public Accountability: Performance Measurement, the Extended State, and the Search for Trust*. Washington, DC: Kettering Foundation.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2005). New Public Management Is Dead--Long Live Digital-Era Governance. *Journal of Public Administration Research and Theory*, *16*(3), 467–494. https://doi.org/10.1093/jopart/mui057

Dunn, W. N., & Fozouni, B. (1976). *Toward a critical administrative theory*. Thousand Oaks, CA: Sage Publications.

Dunn, W. N., & Miller, D. Y. (2007). A Critique of the New Public Management and the Neo-Weberian State: Advancing a Critical Theory of Administrative Reform. *Public Organization Review*, *7*(4), 345–358. https://doi.org/10.1007/s11115-007-0042-3

Durkheim, E. (1897). *Le Suicide*. Paris: F. Alcan.

Durose, C. (2009). Front-line workers and "local knowledge": neighbourhood stories in contemporary UK local governance. *Public Administration*, *87*(1), 35–49. https://doi.org/10.1111/j.1467-9299.2008.01737.x

EBA. (2014). Results - European Banking Authority. Retrieved from https://www.eba.europa.eu/risk-analysis-and-data/eu-wide-stress-testing/2014/results

EBA. (2018). *2018 EU-wide stress test results*. London: European Banking Authority.

EBA. (2019a). EU-wide stress testing. Retrieved from https://eba.europa.eu/risk-analysis-and-data/eu-wide-stress-testing

EBA. (2019b). Supervisory Review and Evaluation Process (SREP) and Pillar 2. Retrieved from https://eba.europa.eu/regulation-and-policy/supervisory-review-and-evaluation-srep-and-pillar-2

Elliott, L. (2016). The EBA's stress tests reveal their own lack of credibility | Business | The Guardian. Retrieved from https://www.theguardian.com/business/economics-blog/2016/aug/01/eba-stress-tests-reveal-their-own-lack-credibility

Enria, A. (2018). *What we have learnt from EU-wide stress tests*. London: European Banking Authority.

Espeland, W. (2015). Narrating numbers. In R. Rottenburg, S. E. Merry, S.-J.

Park, & J. Mugler (Eds.), *The World of Indicators* (pp. 56–75). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316091265.003

Espeland, W. N., & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology*, *113*(1), 1–40. https://doi.org/10.1086/517897

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, *24*, 313–343.

FactMaps. (2016). PISA Worldwide Ranking - average score of math, science and reading. Retrieved from http://factsmaps.com/pisa-worldwide-ranking-average-score-of-math-science-reading/

Fann, K. T. (1970). *Peirce's Theory of Abduction*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-3163-9_1

Farlow, A. (2015). Financial indicators and the global financial crash. In R. Rottenburg, S. E. Merry, S.-J. Park, & J. Mugler (Eds.), *The World of Indicators* (pp. 220–253). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316091265.009

Feron, P. (2013). The impact of quantitative targets on air pollution policies in the Walloon Region (Belgium).

Fioramonti, L. (2013). *Gross domestic problem : the politics behind the world's most powerful number*. London: Zed Books.

Fischer, F., & Forester, J. (1993). *The Argumentative turn in policy analysis*

and planning. Durham, NC: Duke University Press.

Flinders, M. (2014). The Future and Relevance of Accountability Studies. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0004

Flyvbjerg, B. (2001). *Making social science matter : why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

Foucault, M. (1977). *Discipline and punish : the birth of the prison*. London: Allen Lane.

Foucault, M. (2011). *The government of self and others : lectures at the College de France, 1982-1983*. London: Palgrave Macmillan.

Foucault, M., Burchell, G., Gordon, C., & Miller, P. (1991). *The Foucault effect : studies in governmentality : with two lectures by and an interview with Michel Foucault*. Chicago: University of Chicago Press.

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4–5), 663–671. https://doi.org/10.1080/09614520701469955

Gadamer, H.-G. (2004). *Truth and method*. London: Continuum.

Geithner, T. F. (2015). *Stress test : reflections on financial crises*. New York: Penguin Random House.

Giddens, A. (1990). *The consequences of modernity*. Redwood City:

Stanford University Press.

Giddens, A., & Pierson, C. (1998). *Conversations with Anthony Giddens : making sense of modernity*. Stanford: Stanford University Press.

Gillispie, C. C. (1960). *The Edge of Objectivity : an Essay in the History of Scientific Ideas*. Princeton: Princeton University Press.

Goldstein, M. (2015). Bank Stress Tests and Financial Stability: Lessons from the 2009-2014 US and EU-wide tests for Asian emerging economies. In M. Noland & D. Park (Eds.), *From Stress to Growth: Strengthening Asia's Financial Systems in a Post-Crisis World* (pp. 271–316). New York City: Columbia University Press.

Gordon, M. (2009). Toward A Pragmatic Discourse of Constructivism: Reflections on Lessons from Practice. *Educational Studies*, *45*(1), 39–58. https://doi.org/10.1080/00131940802546894

Gore, A. (1993). *From Red Tape to Results: Creating a Government That Works Better & costs Less. Report of the National Performance Review*. Darby, PA: DIANE Publishing.

Gorur, R. (2015). Producing calculable worlds: education at a glance. *Discourse: Studies in the Cultural Politics of Education*, *36*(4), 578–595. https://doi.org/10.1080/01596306.2015.974942

Gorur, R. (2016). Seeing like PISA: A cautionary tale about the performativity of international assessments. *European Educational Research Journal*, *15*(5), 598–616. https://doi.org/10.1177/1474904116658299

Grabosky, P., & Braithwaite, J. (1986). *Of Manners Gentle - Enforcement strategies of Australian business regulatory agencies*. Melbourne: Oxford University Press.

Habermas, Jurgen. (1984). *The theory of communicative action, volume one: Reason and the rationalization of society*. Boston: Beacon Press.

Habermas, Jurgen. (1990). *Moral consciousness and communicative action*. Cambridge, MA: MIT Press.

Habermas, Jurgen, & Derrida, J. (2003). February 15, or What Binds Europeans Together: A Plea for a Common Foreign Policy, Beginning in the Core of Europe. *Constellations*, *10*(3), 291–297. https://doi.org/10.1111/1467-8675.00333

Hackett, E. J. (2008). *The handbook of science and technology studies*. Cambridge, MA: MIT Press.

Hacking, I. (1990). *The taming of chance*. Cambridge, UK: Cambridge Univeristy Press.

Hanegraaff, M., Beyers, J., & De Bruycker, I. (2016). Balancing inside and outside lobbying: The political strategies of lobbyists at global diplomatic conferences. *European Journal of Political Research*, *55*(3), 568–588. https://doi.org/10.1111/1475-6765.12145

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production*

Economics, *154*, 72–80. https://doi.org/10.1016/J.IJPE.2014.04.018

Héritier, A. (2017). Covert Integration of Core State Powers: Renegotiating Incomplete Contracts. In P. Bursens, C. De Landtsheer, L. Braekmans, & B. Segaert (Eds.), *Complex Political Decision-making: Leadership, Legitimacy, and Communication*. Farnham: Ashgate. https://doi.org/10.1093/acprof:oso/9780199662821.003.0012

Hood, C., & Peters, G. (2004). The Middle Aging of New Public Management: Into the Age of Paradox? *Journal of Public Administration Research and Theory*, *14*(3), 267–282. https://doi.org/10.1093/jopart/muh019

Hood, Christopher. (1991). A Public Management for all Seasons. *Public Administration*, *69*(1), 3–19. https://doi.org/10.1111/j.1467-9299.1991.tb00779.x

Hood, Christopher. (2007). Public Service Management by Numbers: Why Does it Vary? Where Has it Come From? What Are the Gaps and the Puzzles? *Public Money and Management*, *27*(2), 95–102. https://doi.org/10.1111/j.1467-9302.2007.00564.x

Hood, Christopher, Rothstein, H., & Baldwin, R. (2001). *The Government of Risk: understanding risk regulation regimes*. Oxford: Oxford University Press. https://doi.org/10.1093/0199243638.001.0001

Hoogenboezem, J. A. (2004). Local government performance indicators in Europe: an exploration. *International Review of Administrative Sciences*, *70*(1), 51–64. https://doi.org/10.1177/0020852304041230

Hopkins, D., Pennock, D., Ritzen, J., Ahtaridou, E., & Zimmer, K. (2008). *External evaluation of the policy impact of PISA*. Paris: OECD.

Hoppe, R. (2009). Scientific advice and public policy: expert advisers' and policymakers' discourses on boundary work. *Poiesis & Praxis*, *6*(3–4), 235–263. https://doi.org/10.1007/s10202-008-0053-3

Horkheimer, M., & Adorno, T. W. (1947). *Dialektik der Aufklärung: Philosophische Fragmente*. Amsterdam: Querido.

Hvidman, U., & Andersen, S. C. (2014). Impact of Performance Management in Public and Private Organizations. *Journal of Public Administration Research and Theory*, *24*(1), 35–58. https://doi.org/10.1093/jopart/mut019

Ingraham, P. W. (2005). Performance: Promises to Keep and Miles to Go. *Public Administration Review*, *65*(4), 390–395. https://doi.org/10.1111/j.1540-6210.2005.00466.x

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377. https://doi.org/10.1016/J.GIQ.2016.08.011

Jasanoff, S. (1986). *Risk Management and Political Culture: A Comparative Study of Science in the Policy Context*. (S. Jasanoff, Ed.) (1st ed.). New York: Russel Sage Foundation.

Jasanoff, S. (2004). *States of knowledge : the co-production of science and social order*. Abingdon-on-Thames: Routledge.

Jasanoff, S. (2006). *Technology as a Site and Object of Politics*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199270439.003.0040

Jessop, B. (2014). Repoliticising depoliticisation: theoretical preliminaries on some responses to the American fiscal and Eurozone debt crises. *Policy & Politics*, *42*(2), 207–223. https://doi.org/10.1332/030557312X655864

Johnston, J. (2004). Assessing government's performance management capability: the case of the Australian electricity industry. *International Review of Administrative Sciences*, *70*(1), 123–136. https://doi.org/10.1177/0020852304041235

Junjan, V. (2015). Strategic Planning in Local Governments in Europe: "Where Do We Go Now"? *Transylvanian Review of Administrative Sciences*, *S.I.*, 45–54.

Kagan, R. A. (1995). What Socio-Legal Scholars Should Do When There Is Too Much Law to Study. *Journal of Law and Society*, *22*(1), 140. https://doi.org/10.2307/1410711

Kaufmann, D., & Kraay, A. (2007). Governance Indicators: Where Are We, Where Should We Be Going? *The World Bank Research Observer*, *23*(1), 1–30. https://doi.org/10.1093/wbro/lkm012

Keane, J. (2009). *The Life and Death of Democracy*. London: Simon & Schuster.

Kelley, J. G., & Simmons, B. A. (2015). Politics by number: Indicators as social pressure in international relations. *American Journal of Political*

Science, *59*(1), 55–70. https://doi.org/10.1111/ajps.12119

Kemper, J., & Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 1–16. https://doi.org/10.1080/1369118X.2018.1477967

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205–223. https://doi.org/10.1177/2053951714528481

Klenk, T., & Reiter, R. (2019). Post-New Public Management: reform ideas and their application in the field of social services: *International Review of Administrative Sciences*, *85*(1), 3–10. https://doi.org/10.1177/0020852318810883

Klüver, H. (2013). Lobbying as a collective enterprise: winners and losers of policy formulation in the European Union. *Journal of European Public Policy*, *20*(1), 59–76. https://doi.org/10.1080/13501763.2012.699661

Kohler-Koch, B. (1996). Catching up with change: The transformation of governance in the European Union. *Journal of European Public Policy*, *3*(3), 359–380. https://doi.org/10.1080/13501769608407039

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kvale, S., & Brinkmann, S. (1996). *Interviews: an introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage Publications.

Latour, B. (1987). *Science in action: How to follow scientists and engineers*

through society. Harvard University Press.

Latour, B. (1988a). Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer* Reinventing the Door. *Social Problems*, *35*(3), 298–310. Retrieved from https://www.nyu.edu/projects/nissenbaum/papers/Latour_Mixing.pdf

Latour, B. (1988b). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

Latour, B. (1991). *We have never been modern*. Cambridge, Massachusetts: Harvard University Press.

Latour, B. (1999). On recalling ANT. *The Sociological Review*, *47*(1), 15–25. https://doi.org/10.1111/j.1467-954X.1999.tb03480.x

Latour, B. (2005). *Reassembling the social : an introduction to actor-network-theory* (1st ed.). Oxford: Oxford University Press.

Latour, B. (2013). *An inquiry into modes of existence: An anthropology of the moderns*. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life : the construction of scientific facts*. Princeton: Princeton University Press.

Law, J. (1999). After Ant: Complexity, Naming and Topology. *The Sociological Review*, *47*(1), 1–14. https://doi.org/10.1111/j.1467-954X.1999.tb03479.x

Law, J. (2004). *After Method: Mess in Social Science Research*. Abingdon-on-Thames: Routledge.

Lemke, T. (2011). *Foucault, governmentality, and critique*. Colorado: Paradigm Publishers.

Levi-Faur, D. (2005). The Global Diffusion of Regulatory Capitalism. *The Annals of the American Academy of Political and Social Science*, *598*(1), 12–32. https://doi.org/10.1177/0002716204272371

Levitsky, S., & Lucan, A. W. (2009). Why Democracy Needs a Level Playing Field. *Journal of Democracy*, *21*(1), 57–68. https://doi.org/10.1353/jod.0.0148

Lilleker, D. G. (2003). Interviewing the Political Elite: Navigating a Potential Minefield. *Politics*, *23*(3), 207–214. https://doi.org/10.1111/1467-9256.00198

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks: Sage Publications.

Lincoln, Y. S., & Guba, E. G. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denkzin & Y. S. Lincoln (Eds.), *The handbook of qualitative research* (pp. 163–188). Beverly Hills, CA: Sage Publications.

Lindner, P. (2017). Reset Modernity!: Re-Present, Re-Set, Re-Assemble—Bruno Latour |ZKM| Center for Arts and Media Karlsruhe, 16 April–21 August 2016. *GeoHumanities*, *3*(1), 209–217. https://doi.org/10.1080/2373566X.2016.1259007

Lodge, M., & Wegrich, K. (2012). *Managing regulation : regulatory analysis, politics and policy*. London: Palgrave Macmillan.

Lowery, D. (2013). Lobbying influence: Meaning, measurement and missing. *Interest Groups & Advocacy*, *2*(1), 1–26. https://doi.org/10.1057/iga.2012.20

Lyotard, J.-F., & Massumi, B. (1984). *The postmodern condition : a report on knowledge* (1st ed.). Minnesota: University of Minnesota Press.

MacKenzie, D. A. (2006). *An engine, not a camera : how financial models shape markets*. Cambridge, MA: MIT Press.

Mahmood, M., Weerakkody, V., & Chen, W. (2019). The role of information and communications technology in the transformation of government and citizen trust: *International Review of Administrative Sciences*. https://doi.org/10.1177/0020852318816798

Majone, G. (1994). The rise of the regulatory state in Europe. *West European Politics*, *17*(3), 77–101. https://doi.org/10.1080/01402389408425031

Manovich, L. (2011). Trending: the promises and the challenges of big social data. In M. Gold & K. Minneapolis (Eds.), *Debates in the Digital Humanities*. Minnesota: The University of Minnesota Press.

Marcuse, H. (1941). *Reason and Revolution: Hegel and the rise of social theory*. Oxford: Oxford University Press.

Marlier, E., & Atkinson, A. B. (2010). Indicators of poverty and social exclusion in a global context. *Journal of Policy Analysis and Management*, *29*(2), 285–304. https://doi.org/10.1002/pam.20492

Martin, K. (2018). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *12*(2), 1–16. https://doi.org/10.1007/s10551-018-3921-3

Matheus, R., Janssen, M., & Maheshwari, D. (2018). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*. https://doi.org/10.1016/J.GIQ.2018.01.006

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data : a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

McCarthy, T. (1990). The Critique of Impure Reason. *Political Theory*, *18*(3), 437–469. https://doi.org/10.1177/0090591790018003005

McGlynn, E. A., & Asch, S. M. (1998). Developing a Clinical Performance Measure. *American Journal of Preventive Medicine*, *14*(3), 14–21. https://doi.org/10.1016/S0749-3797(97)00032-9

McLaughlin, K., Osborne, S. P., & Ferlie, E. (2002). *New public management : current trends and future prospects*. Abingdon-on-Thames: Routledge.

Merton, R. K. (1968). *Social theory and social structure*. New York: The Free Press.

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, *367*(16), 1562–1564. https://doi.org/10.1056/NEJMon1211064

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, *83*(2), 340–363. https://doi.org/10.1086/226550

Micheli, P., & Pavlov, A. (2017). What is performance measurement for? Multiple uses of performance information within organizations. *Public Administration*. https://doi.org/10.1111/padm.12382

Mikuła, Ł., & Kaczmarek, U. (2019). From marketization to recentralization: the health-care system reforms in Poland and the post-New Public Management concept. *International Review of Administrative Sciences*, *85*(1), 28–44. https://doi.org/10.1177/0020852318773429

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis : an expanded sourcebook* (2nd ed.). Beverly Hills, CA: Sage.

Miller, G. J. (1992). *Managerial dilemmas : the political economy of hierarchy*. Cambridge University Press.

Miller, P. (2004). Governing by numbers: why calculative practices matter. In A. Amin & N. J. Thrift (Eds.), *The Blackwell cultural economy reader. Blackwell readers in geography* (pp. 179–190). Malden: Blackwell.

Mills, C. W. (2000). *The sociological imagination*. Oxford: Oxford University Press.

Mitchell, T. (1988). *Colonising Egypt*. University of California Press. Retrieved from http://library1.org/_ads/94BB91E6296EE5A553DD0E9D76B0078B

Mol, N. P., & De Kruijf, J. A. M. (2004). Performance management in Dutch central government. *International Review of Administrative Sciences*, *70*(1), 33–50. https://doi.org/10.1177/0020852304041229

Mouzelis, N. P. (1967). *Organisation and bureaucracy : an analysis of modern theories*. Chicago, IL: Aldine.

Moynihan, D. P. (2008). *The dynamics of performance management : constructing information and reform*. Washington: Georgetown University Press.

Moynihan, D. P., & Kroll, A. (2016). Performance Management Routines That Work? An Early Assessment of the GPRA Modernization Act. *Public Administration Review*, *76*(2), 314–323. https://doi.org/10.1111/puar.12434

Mügge, D. (2016). Studying macroeconomic indicators as powerful ideas. *Journal of European Public Policy*, *23*(3), 410–427. https://doi.org/10.1080/13501763.2015.1115537

Mulgan, R. (2003). *Holding Power to Account: Accountability in Modern Democracies*. New York: Palgrave.

Nielsen, P. A. (2014). Performance Management, Managerial Authority, and Public Service Performance. *Journal of Public Administration Research and Theory*, *24*(2), 431–458. https://doi.org/10.1093/jopart/mut025

Niemeijer, D. (2002). Developing indicators for environmental policy: data-driven and theory-driven approaches examined by example. *Environmental Science & Policy*, *5*(2), 91–103.

https://doi.org/10.1016/S1462-9011(02)00026-6

O'Halloran, S., Maskey, S., McAllister, G., Park, D. K., & Chen, K. (2015). Big Data and the Regulation of Financial Markets. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15* (pp. 1118–1124). New York, New York, USA: ACM Press. https://doi.org/10.1145/2808797.2808841

O'Malley, M. (2014). Doing What Works: Governing in the Age of Big Data. *Public Administration Review*, *74*(5), 555–556. https://doi.org/10.1111/puar.12260

O'Neil, C. (2016). *Weapons of math destruction : how big data increases inequality and threatens democracy*. New York: Crown.

O'Neill, O. (2002). *A question of trust*. New York: Cambridge University Press.

OECD. (2019). What is PISA? Retrieved from http://www.oecd.org/pisa/

Osborne, D., & Gaebler, T. (1992). *Reinventing government : how the entrepreneurial spirit is transforming the public sector*. New York: Plume.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Patterson, M. (1998). Commensuration and theories of value in ecological

economics. *Ecological Economics*, *25*(1), 105–125. https://doi.org/10.1016/S0921-8009(97)00166-3

Peeters, H., & Verschraegen, G. (2013). Governance by numbers: risico's verbonden aan de internationale benchmarking en ranking van pensioensystemen *. *Beleid En Maatschappij*, *40*(2), 133–155.

Peeters, H., Verschraegen, G., & Debels, A. (2014). Commensuration and policy comparison: How the use of standardized indicators affects the rankings of pension systems. *Journal of European Social Policy*, *24*(1), 19–38. https://doi.org/10.1177/0958928713511279

Pidd, M. (2005). Perversity in public service performance measurement. *International Journal of Productivity and Performance Management*, *54*(5/6), 482–493. https://doi.org/10.1108/17410400510604601

Pires, R. R. C. (2011). Beyond the fear of discretion: Flexibility, performance, and accountability in the management of regulatory bureaucracies. *Regulation & Governance*, *5*(1), 43–69. https://doi.org/10.1111/j.1748-5991.2010.01083.x

Polanyi, M., & Nye, M. J. (1962). *Personal knowledge : towards a post-critical philosophy*. New York: Harper Torch Books.

Pollack, M. A. (2003). *The engines of European integration: delegation, agency, and agenda-setting in the EU*. Oxford: Oxford University Press.

Pollitt, C. (2018). Performance management 40 years on: a review. Some key decisions and consequences. *Public Money & Management*, *38*(3), 167–174. https://doi.org/10.1080/09540962.2017.1407129

Pollitt, C., & Talbot, C. (2004). *Unbundled Government: A Critical Analysis of the Global Trend to Agencies*. London: Routledge.

Porter, T. M. (1995). *Trust in numbers : the pursuit of objectivity in science and public life*. Princeton: Princeton University Press.

Porter, T. M. (2015). The flight of the indicator. In R. Rottenburg, S. E. Merry, S.-J. Park, & J. Mugler (Eds.), *The World of Indicators* (pp. 34–55). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316091265.002

Power, M. (1997). *The audit society : rituals of verification*. Oxford: Oxford University Press.

Power, M. (2003). Auditing and the production of legitimacy. *Accounting, Organizations and Society*, *28*(4), 379–394. https://doi.org/10.1016/S0361-3682(01)00047-2

Quagliariello, M. (2009). *Stress-testing the Banking System*. (M. Quagliariello, Ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511635618

Rabin, J., & Bowman, J. S. (1984). *Politics and administration : Woodrow Wilson and American public administration*. New York: Dekker. Retrieved from https://www.worldcat.org/title/politics-and-administration-woodrow-wilson-and-american-public-administration/oclc/10022785

Radin, B. A. (2006). *Challenging the performance movement : accountability, complexity, and democratic values*. Washington, DC:

Georgetown University Press.

Rafailov, D. (2011). The Failures of Credit Rating Agencies during the Global Financial Crisis - Causes and Possible Solutions. *Economic Alternatives*, (1), 34–45.

Ravetz, J. R. (1971). *Scientific knowledge and its social problems*. Piscataway, NJ: Transaction Publishers.

Reiter, R., & Klenk, T. (2018). The manifold meanings of 'post-New Public Management' – a systematic literature review: *International Review of Administrative Sciences*, *85*(1), 11–27. https://doi.org/10.1177/0020852318759736

Renou, Y. (2017). Performance indicators and the new governmentality of water utilities in France. *International Review of Administrative Sciences*, *83*(2), 378–396. https://doi.org/10.1177/0020852315589696

Roness, P. G., Verhoest, K., Rubecksen, K., & MacCarthaigh, M. (2008). Autonomy and Regulation of State Agencies: Reinforcement, Indifference or Compensation? *Public Organization Review*, *8*(2), 155–174. https://doi.org/10.1007/s11115-008-0057-4

Rose, N., O'Malley, P., & Valverde, M. (2006). Governmentality. *Annual Review of Law and Social Science*, *2*(1), 83–104. https://doi.org/10.1146/annurev.lawsocsci.2.081805.105900

Rottenburg, R., & Merry, S. E. (2015). A world of indicators: The making of governmental knowledge through quantification. In R. Rottenburg, S. E. Merry, S.-J. Park, & J. Mugler (Eds.), *The World of Indicators* (pp. 1–

33). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316091265.001

Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton: Princeton University Press.

Sarfaty, G. (2011). Regulating Through Numbers: A Case Study of Corporate Sustainability Reporting. *Virginia Journal of International Law*, *53*, 575–624. https://doi.org/http://dx.doi.org/10.2139/ssrn.1915212

Schaffer, S. (1989). Glass Works: Newton's Prisms and the Uses of Experiment. In S. Gooding, David; Pinch, Trevor; Schaffer (Ed.), *The Uses of Experiment: Studies in the Natural Sciences* (pp. 67–89). Cambridge: Cambridge University Press.

Schmidt, V. H. (2001). Oversocialised Epistemology: A Critical Appraisal of Constructivism. *Sociology*, *35*(1), 135–157. https://doi.org/10.1177/0038038501035001009

Schneibel, G., & Braun, M. (2010, December 13). Stress tests doubted as Ireland suffers from bank bailout | Business| Economy and finance news from a German perspective. *DW*. Retrieved from https://www.dw.com/en/stress-tests-doubted-as-ireland-suffers-from-bank-bailout/a-6321854

Schwandt, T. A. (2001). Understanding Dialogue as Practice. *Evaluation*, *7*(2), 228–237. https://doi.org/10.1177/13563890122209658

Schwartz-Shea, P., & Yanow, D. (2012). *Interpretive research design :*

concepts and processes. Abingdon-on-Thames: Routledge.

Scott, J. C. (1998). *Seeing like a state : how certain schemes to improve the human condition have failed*. New Haven, Connecticut: Yale University Press.

Scott, J. C., & Trubek, D. M. (2002). Mind the Gap: Law and New Approaches to Governance in the European Union. *European Law Journal*, *8*(1), 303–326.

Smith, P. (1995). Performance indicators and outcome in the public sector. *Public Money & Management*, *15*(4), 13–16. https://doi.org/10.1080/09540969509387889

Smolan, R., & Erwitt, J. (2012). *The human face of big data*. New York: Against All Odds Productions.

Smulewicz-Zucker, G. (2017). The Frankfurt School and the Critique of Instrumental Reason. In M. Thopson (Ed.), *The Palgrave Handbook of Critical Theory. Political Philosophy and Public Purpose*. (pp. 185–206). New York: Palgrave Macmillan.

Stebbins, R. A. (2001). *Exploratory research in the social sciences*. Thousand Oaks, CA: Sage Publications.

Stone, C. (2012). Problems of Power in the Design of Indicators of Safety and Justice in the Global South. In K. E. Davis, A. Fisher, B. Kingsbury, & S. E. Merry (Eds.), *Governance by indicators: global power through quantification and ranking* (pp. 281–294). Oxford: Oxford University Press.

Stone, D. A. (2002). *Policy paradox : the art of political decision making*. New York: Norton.

Taylor, J. (2011). Factors influencing the use of performance information for decision making in Australian state agencies. *Public Administration*, *89*(4), 1316–1334. https://doi.org/10.1111/j.1467-9299.2011.02008.x

Termeer, C. J. a. M., Dewulf, A., Breeman, G., & Stiller, S. J. (2013). Governance Capabilities for Dealing Wisely With Wicked Problems. *Administration & Society*, *47*(6), 680–710. https://doi.org/10.1177/0095399712469195

Tervonen-Gonçalves, L. (2012). From averages to best performers: use of comparisons in identity formation. *Critical Policy Studies*, *6*(3), 304–323. https://doi.org/10.1080/19460171.2012.717784

Thedvall, R. (2012). Negotiating impartial indicators: putting transparency into practice in the EU. *Journal of the Royal Anthropological Institute*, *18*(2), 311–329. https://doi.org/10.1111/j.1467-9655.2012.01745.x

Thun, C. (2013). Are Regulatory Stress Tests Cost Without Value? Retrieved from https://www.moodysanalytics.com/risk-perspectives-magazine/stress-testing-europe/rethinking-stress-testing/are-regulatory-stress-tests-just-cost-without-value

Timmermans, S., & Tavory, I. (2012). Theory Construction in Qualitative Research. *Sociological Theory*, *30*(3), 167–186. https://doi.org/10.1177/0735275112457914

Tissot, B. (2017). *Big data and central banking: Overview of the IFC satellite*

meeting. Retrieved from https://ssrn.com/abstract=2577759.

Titcomb, J. (2014). Eurozone banks rush to fill black hole ahead of stress tests - Telegraph. Retrieved from https://www.telegraph.co.uk/finance/newsbysector/banksandfinance/11155339/Eurozone-banks-rush-to-fill-black-hole-ahead-of-stress-tests.html

Turner, S. (2008). The Social Study of Science before Kuhn. In E. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies* (pp. 33–62). California: The MIT Press.

van der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, *36*(1), 27–38. https://doi.org/10.1016/J.GIQ.2018.10.011

van Dijk, T. A. (2008). *Discourse and power*. London: Palgrave Macmillan.

Van Dooren, W., Bouckaert, G., & Halligan, J. (2015). *Performance Management in the Public Sector*. London: Routledge.

Van Dooren, W., & Hoffmann, C. (2018). Performance management in Europe: An idea whose time has come - and gone? In E. Ongaro & S. VanThiel (Eds.), *The Palgrave Handbook of Public Administration and Management in Europe* (pp. 207–225). London: Palgrave Macmillan. https://doi.org/10.31235/OSF.IO/9ZYV8

van Hoek, R. I. (1998). "Measuring the unmeasurable" - measuring and

improving performance in the supply chain. *Supply Chain Management: An International Journal*, *3*(4), 187–192. https://doi.org/10.1108/13598549810244232

van Ostaijen, M., & Scholten, P. (2017). The politics of numbers. Framing intra-EU migrants in the Netherlands. *Critical Policy Studies*, *11*(4), 477–498. https://doi.org/10.1080/19460171.2016.1224725

van Steen, M. (2015). Regulatory Big Data: Goals and Initiatives. Retrieved from https://www.moodysanalytics.com/risk-perspectives-magazine/risk-data-management/regulatory-spotlight/regulatory-big-data-regulator-goals-and-global-initiatives

van Thiel, S., & Leeuw, F. L. (2002). The Performance Paradox in the Public Sector. *Public Performance & Management Review*, *25*(3), 267–281. https://doi.org/10.1080/15309576.2002.11643661

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology*, 206–224. https://doi.org/10.1080/13600869.2017.1298547

Vickerman, R. (2007). Cost — Benefit Analysis and Large-Scale Infrastructure Projects: State of the Art and Challenges. *Environment and Planning B: Planning and Design*, *34*(4), 598–610. https://doi.org/10.1068/b32112

Wagenaar, H. (2011). *Meaning in action : interpretation and dialogue in policy analysis*. Armonk: M.E. Sharpe.

Walker, R. M., Damanpour, F., & Devece, C. A. (2011). Management

Innovation and Organizational Performance: The Mediating Effect of Performance Management. *Journal of Public Administration Research and Theory*, *21*(2), 367–386. https://doi.org/10.1093/jopart/muq043

Walshe, K., Harvey, G., & Jas, P. (2010). *Connecting Knowledge and Performance in Public Services: From Knowing to Doing*. Cambridge: Cambridge Univeristy Press.

Weber, M. (1919). *Science as a Vocation*. Munich: Duncker & Humblodt.

Welsh, J. (2017). Ranking academics: toward a critical politics of academic rankings. *Critical Policy Studies*, 1–21. https://doi.org/10.1080/19460171.2017.1398673

Winkler, B. (2000). Which Kind of Transparency? On the Need for Clarity in Monetary Policy-Making. *ECB Working Paper*, *26*, 35.

Woll, C. (2014). *The power of inaction : bank bailouts in comparison*. Ithaca: Cornell University Press.

Yanow, D., & Schwartz-Shea, P. (2006). *Interpretation and method : empirical research methods and the interpretive turn*. Armonk: M.E. Sharpe.

Yin, R. K. (2012). *Applications of case study research*. Thousand Oaks: SAGE Publications.

Zifcak, S. (1994). *New managerialism : administrative reform in Whitehall and Canberra*. London: Open University Press.

APPENDICES

APPENDIX 1: TOPIC GUIDE (BANKS)

1. Introduction Introduce myself (affiliation, funding,…) Introduce topic Explain aims Explain confidentiality and anonymity Check duration of interview Check if they have any other questions
2. Background 'Warm respondent up' + get background info: Find out how respondent became, and is, involved in the stress testing procedure. What their position and expertise is. - What do they do? - What did they do before this? - How long involved in STing
3. "just a communication exercise" How 'accurate' is the ST? Correct representation of bank's risk vs. 'just a story'/'communication tool' or both? (also notion of 'factity' + 'gaming') - Tell me about stress testing - Where did ST come from (why did banks/regulators) start

- "Do you remember what you thought about ST when the exercises just started?"
- Is it credible, why?
- How mechanical is the ST (vs how much 'grey zone' is there)
- How is it used by banks as a communication tool?
- How is it used by ECB as communication tool?
- What are strengths, weaknesses
- What are other measurement tools
- (good that it's used in SREP?)

4. "Level playing field"

Is there a LPF, why (not)? How is the LPF put into practice? How do you understand LPF?

- Any difficulties conducting the ST, why? [ask for examples]
    - Probe: what do you struggle with and why?
- Do they feel there is a LPF (how could one be put in place)
- discuss balance between level playing field (LPF) and 'true' representation of banks

5. "Lessons learned"

What have different actors 'learned' about banks from the stress test (does ST clarify bank's health – for who? does it make financial sector more transparent?), what's the value, is there value?

- Does respondent learn about own bank (ito risk) from ST
- Does respondent think other, external actors (board of directors, policy makers, analysts…) learn more about risk of bank?
- ST added to long list of existing indicators, is it really added value, why?

- Did ST results come as a 'surprise' for you, for others in the bank?
- (link back to 'accuracy') Is learning in contrast with making bank 'look good'? (both for bank & ECB)

6. "Change/impact"

What is the (regulatory) impact of the ST? (e.g. data transparency, input SREP, reputational damage…) Have things changed since results?

- What is the impact of the ST on your bank?
- How is the role of the ECB and the banks evolving in the ST?
- How do you feel about this?
- What are implications for power& accountability?

7. Cool down

Winding down, make sure respondent will leave interview satisfied.

- Anything we haven't covered yet
- Any final remarks
- Any questions
- Ask for other possible respondents

8. In conclusion

Make necessary practical arrangements. Express gratitude, reiterate confidentiality, discuss how respondent might be quoted

- Thank for time, interest, effort
- Reminder confidentiality
- Can I quote you on what you've said today I would you like to go over the final quotes I end up using?
- (Arrange next meeting)

Appendix 2: List of interviews

Respondent	Date interview
Round 1: Banks and consulting firms (2015/2016)	
Bank A respondent 1	3/12/15
Bank A respondent 1	8/8/16
Bank A respondent 2	8/8/16
Bank B respondent 1	23/12/15
Bank B respondent 1	25/2/16
Bank B respondent 1	25/6/16
Bank B respondent 2	25/2/16
Bank B respondent 2	27/6/16
Bank C respondent 1	18/12/15
Bank C respondent 2	7/11/16
Bank D respondent 1	27/4/15
Bank D respondent 1	14/12/15
Bank D respondent 2	19/5/15
Bank D respondent 2	22/2/16
Bank D respondent 2	10/10/16
Bank D respondent 3	22/2/16
Bank D respondent 4	7/3/16
Bank D respondent 5	29/7/16
CONSULTING A respondent 1	17/6/15
CONSULTING A respondent 2	17/6/15
CONSULTING A respondent 2	17/10/16
CONSULTING B respondent 1	3/5/16
Round 2: ECB (2017)	
ECB 1	7/3/17
ECB 2	8/3/17
ECB 3	9/3/17
ECB 4	10/3/17

ECB 5	14/3/17
ECB 6	21/3/17
ECB 7	22/3/17
ECB 8	27/3/17
Round 3: banks, consulting firms, EBA, NBB (2018)	
Bank A respondent 1	20/2/18
Bank A respondent 2	20/2/18
Bank A respondent 3	20/2/18
Bank A respondent 4	20/2/18
Bank A respondent 5	20/2/18
Bank B respondent 2	2/3/18
Bank C respondent 3	16/2/18
Bank C respondent 4	16/2/18
Bank D respondent 2	7/3/18
CONSULTING C respondent 1	6/4/18
EBA 1	16/4/18
EBA 2	16/4/18
EBA 3	16/4/18
NBB 1	15/6/18
NBB 2	15/6/18

APPENDIX 3: CODE BOOK

Initial codes (emergent coding)	Sensitizing concepts
Methodological constraints	Calculating stress test
Granular data requirements	
Top down vs. bottom up	
Level playing field (LPF)	
QA process	
Gaming	
National context	
Resource intensive	
Automatization	
Internal models	
Unclear results (do not understand results)	Results of stress test
Unrealistic results	
Fair results	
Political results	
Weak results	
Reassure markets (communication ex)	Purpose / use of stress test
Thorough check banks	
Use in SREP	
Compare EU banks	
Reinstate trust wider public (comm. ex)	
Understand risk	
Get granular data banks	
Microprudential vs macroprudential policy	
More money risk departments	Changes in bank due to stress test
No changes	
Cooperation across departments	
Learning opportunities	
IT investments	
Changes asset classes	

In the aftermath of the 2008 financial crisis, governments introduced stress tests to measure and monitor banks' health. 'Breaking the Bank' assesses whether the EU-wide banking stress test is a good regulatory tool.

It deals with some of the key questions that still linger about the exercise. Do the benefits of the stress test outweigh the enormous cost? It is tough enough on banks? And does it tell us which banks are healthy or not?

Beyond the scope of the stress test, the book addresses how indicators are made, and to what effect. This is increasingly important as indicators are produced through socio-political power plays and negotiations. Before we can use these indicators in policy and public debate, we need to understand where they come from and what they do.

Proefschrift voorgelegd tot het behalen van de graad van doctor in de Sociale Wetenschappen: Politieke Wetenschappen aan de Universiteit Antwerpen. Te verdedigen door Shirley KEMPENEER.

Faculty of Social Sciences, Antwerp 2019
promotor prof. dr. Wouter Van Dooren