AARHUS BSS
SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Contemporary performance measurement and causal thinking

PhD dissertation

Kristian Mohr Røge

Aarhus BSS
Aarhus University
Department of Management
2017

ACKNOWLEDGEMENTS

As I was about to start on my PhD one of my supervisors told me in a tone of sarcasm 'welcome to three years in hell'. In some sense, he was right, as the process of writing a PhD is a long journey filled with periods of frustrations and despair. It is without doubt a very challenging task to be writing on a project for three years, but it is also enlightening and a privilege. However, when I now look back at the three years, which has passed, I must say that I for the most parts have enjoyed my time as a PhD student. I largely attribute this to the people surrounding me, who all in their own way have contributed in making my journey enjoyable.

In writing this dissertation, I have been given invaluable feedback, comments, and encouragements from colleagues, fellow researchers, friends and family. First, it has been a great pleasure to write this dissertation under the supervision of Hanne Nørreklit and Morten Jakobsen. Hanne Nørreklit is truly one of the most innovative and visionary researchers in the field of management accounting and for me it has been a great pleasure and opportunity to work with her and our discussions and collaborations have inspired many aspects of my work. Morten Jakobsen has for me been an inspiration through his integrity in all aspects of his work and I hope his influence will follow me throughout my own career. Additionally, I would give a special thanks to Niels Joseph Lennon who not only have been a great colleague, but also been a strong support at times when the PhD was mainly frustrations and despair. I would also like to thank Nikolaj Kure whom I have been working with on a few research projects. Hanne, Morten, Niels and Nikolaj, I hope you have enjoyed working together as much as I have.

Lastly, I would like to thank Rafael Heinzelmann, who welcomed me in Bergen and I have enjoyed our many discussions and greatly appreciated the feedback I received during my stay at the Norwegian School of Economics (NHH). It was a great experience for me to spend the spring of 2016 at NHH in Bergen. I would also like to thank the managers of the 'municipality'. Without your collaboration, this would not have been possible.

Lastly, my thanks go to people outside academia – friends and family who have been there along the process. Thanks to my daughter Naomi for reminding me about life outside the PhD. Last but certainly not least, thanks to my beloved wife, Durita. Thanks for believing in me when times were darkest and for your continuous encouragement and support no matter what.

Kristian Mohr Røge, October 2017.

STATUS OF THE PAPERS

The dissertation includes four papers. This is the status of the four papers:

Chapter 2: 'Causality in contemporary performance measurement: Are causal questions being answered?'.

Chapter 3: 'Is the validity of positivistic management accounting research exposed to questionable research practices?' The paper was presented at the 10th conference for new directions in management accounting in Brussel. The paper has been rewritten, and is currently under review in Accounting, Organization and Society.

Chapter 4: 'A study on the criteria of internal transparency, efficiency and effectiveness in measuring local government performance'. The paper was presented at NPS 2016 'Transparency and Trust in Public Services' special issue conference at Edinburgh Business School. The paper is currently in second round *revise and re-submit* in Financial Accountability & Management. A Danish version of the paper has been published in Økonomistyring og Informatik.

Chapter 5: 'The illusion of 'objective and result-based management': Beyond a NPM tool in Denmark' The paper was presented at the 7th conference on actor-reality construction at Tampere University of Technology. The paper has been submitted to British Accounting Review.

EXECUTIVE SUMMARY

Traditionally, performance measurement is about quantifying the efficiency and effectiveness of organisational actions and performance measures are therefore metrics used to *quantify* the efficiency and effectiveness of actions. This is a narrow definition, which is also why it does not convey the complete label of contemporary performance measurement. Performance measurement is now often considered a multi-dimensional system that incorporates both financial as well as non-financial measures, it includes both internal and external measures of performance, which quantify what has been achieved and predicts the future. Performance measurement is not done in isolation, as it is only relevant within a framework against which the efficiency and effectiveness of an action can be judged. This frame is typically the strategy of the organisation from which the performance measures must be developed. Performance measurement also influences the environment in which it operates. When starting to measure, deciding what to measure, how to measure and what the targets should be, are all actions that influence the individuals within the organisation. This means that once measurement has started, the performance reviews will inevitable have consequences, which is also why performance measurement is considered an integral part of the strategic planning and control system of the organisation being measured.

Performance measurement continues to be in an evolutionary process. However, when evolution leads to postulates of generic strategic actions, which are claimed to drive future successful financial performance and make prescriptions for managerial actions founded on the presumption of causality between non-financial measures and financial performance, it becomes a 'controversial' evolution. The idea of causality initially received critique, but has since remained persistent in theory and practice while serving as a foundation for the importance of non-financial measures in performance measurement theory. It also provides performance measurement with a future orientation and is a feature long desired for in the late 1980s and early 1990s. On the other hand, if non-financial measures were found to be *not* causally linked to organisational performance, it could bring their relevance into question.

The dissertation aims at contributing to contemporary performance measurement theory in two ways. First, we analyse the empirical grounding of causality, and second, we study public sector implementations of contemporary performance measurement. A discussion of causality is an area which has remained largely neglected in performance measurement theory and by studying causality from these two distinct ways, we intend to overcome the, often, one-sided perception of causality and instead develop a thorough understanding of the role causality plays in contemporary performance measurement and the validity of the presumption. The message in this dissertation is that we possess very little evidence for the existence of true universal causations that can guide practice through postulates of generic strategic actions that are to ensure

successful business performance and hence make prescriptions for managerial actions. The dissertation consists of four papers and contributes as follows:

The first paper search published positivistic management accounting research (PMAR) for empirical evidence for the existence of causality in contemporary performance measurement literature. It analyses the progress of PMAR on uncovering consistent causal relationships between non-financial measures and financial performance and, as such, we provide clarity on the validity of the presumption of causality. In conclusion, we find that the empirical ground for claiming the existence of causality between non-financial measures and financial performance is at best weak, as there is a clear lack of consistency in the empirical results. We are therefore unable to provide the empirical evidence needed for justifying the transformation of causality from a brute fact to a stylized fact.

The second paper is analysing if the publication practices of PMAR are allowing for the phenomenon of questionable research practices (QRPs) to take root. The consequences of QRPs are a distortion of the hypothetico-deductive method in favour of a researchers' own hypothesis with the side-effect of increasing the likelihood of experiencing a false-positive. QRPs have been found to be widespread in natural and social science. If the publication practices of positivistic management accounting research are unintentionally incentivising QRPs, we would expect QRPs to be present in PMAR. We find that the current publication practices of PMAR provide space for QRPs to flourish, and we would therefore expect that the ratio of false-positives in PMAR is well above the assumed ratio of 5 percent. In consequence, we question if the advocacy of causality between non-financial measures and financial performance is reasonable considering the inconsistency of empirical evidence and that the ratio of false-positives is expectedly higher than the conventional ratio of 5 percent.

The third paper studies the use of contemporary performance measurement in the Danish public sector. It explores how the efficiency and effectiveness criteria relate to the inadequacy of performance measurement implementations in the public sector. The paper finds, that notwithstanding the endeavours to develop a well-functioning and successful performance measurement system (PMS), the analysis argues that the PMS fails to accomplish its purpose of directing, actions and activities toward the achievement of strategic objectives. As it ends up being a PMS that is unable to create internal transparency, and therefore efficiency and effectiveness cannot be balanced through the activities of management control. The PMS become an administrative burden that provides top management with a false sense of security in the optimisation of scarce resources.

The fourth paper is a case study on the validity of the PMS and NPM tool 'objective and result-based management' developed and propagated by the Agency of Modernisation. It is a study on how performance measurement and management models in themselves may contain deficiencies that create problems for validity and how these issues unfold in practice. The study

addresses a need for a better understanding of why the implementations of such new public management tools appear to be continuously failing in the Danish public sector. Our analysis evidences that 'objective and result-based management' is poorly outlined and with mismatches in its conceptual structure that leads to a language game of illusions. These issues are then translated into the practical application of performance contracts between the Ministry of Higher Education and Science and the Danish universities. An objective of the contracts was to increase educational quality through a schematic of push causalities, but it was not possible to grasp what perception of quality the universities were striving towards or towards what type of quality the ministry wished them to strive. This resulted in a highly diverse formulation of measures that appeared to be based more on impressions than rational reasoning, which rendered the content open for interpretation; in other words, we have no clue whether the university educations are of an increasing quality or high quality if we solely look at the measures in the contract. In the end, we found the outlined performance management framework of 'objective and result-based management' not to live up to the basic principles for providing concepts that can facilitate the purpose of creating effective public sector institutions.

DANSK RESUMÉ

Traditionelt set handler præstationsmåling om, at måle efficiensen og effektiviteten af organisatoriske handlinger. Præstationsmål er dermed måltal der bruges til at *kvantificere* efficiensen og effektiviteten af handlinger. Dette er en snæver definition af præstationsmåling, hvilket også er årsagen, til at den ikke længere udfylder den moderne forståelse af præstationsmåling. Præstationsmåling bliver nu betragtet som et multidimensionalt system, der indeholder finansielle og ikke-finansielle måltal, samt interne og eksterne måltal, som kan kvantificere, hvad der er blevet opnået, og hvad der kommer til at ske i fremtiden. Præstationsmåling kan ikke ske i isolation, da der skal være en referenceramme at måle efficiens og effektivitet op imod. Denne ramme er typisk den organisatoriske strategi, fra hvilken målene er blevet formuleret. Præstationsmåling influerer desuden det miljø det opererer i. Hvilket betyder at når en organisation vælger at arbejde med præstationsmåling, så er beslutningen om hvad der skal måles, hvordan det skal måles, og hvad målsætningerne er, alle beslutninger som påvirker aktørerne i organisationen. Dermed vil præstationsmåling uundgåeligt medføre en lang række konsekvenser, utilsigtede som tilsigtede, hvilket også er begrundelsen for hvorfor præstationsmåling betragtes som en central del af den strategiske planlægning og kontrolsystemet i en organisation.

Præstationsmåling er et evolutionært begreb, men når det hævdes, at postulater om generiske strategiske handlinger og forskrifter for ledelseshandlinger er drivkraften bag en fremtidig succesfuld organisatorisk præstation, baseret på forestillingen om kausalitet mellem ikke-finansielle og finansielle præstationer, så bliver det en 'kontroversiel' evolution. Kausalitetsantagelsen blev oprindeligt skeptisk modtaget, men har sidenhen fået cementeret sig i præstationsmålingsteorien, og det er blevet fundamentet for betydningen af ikke-finansielle måltal. Kausalitet og ikke-finansielle måltal leverer en orientering mod fremtiden, hvilket var en egenskab, som var højt ønsket i slutningen af 80'erne og start 90'erne. På den anden side, hvis ikke-finansielle måltal findes at være urelateret til den organisatoriske præstation vil deres relevans være betvivlet.

Denne afhandling har til formål at bidrage til moderne præstationsmåling på to måder. For det første ved at analysere argumentet bag antagelsen om kausalitet, og for det andet ved at studere og analysere to forskellige offentlig sektor cases, hvor moderne præstationsmåling er blevet implementeret. Afhandlingen er dermed et forsøg på at opnå en mere detaljeret forståelse af den rolle kausalitet spiller i moderne præstationsmåling og den teoretiske gyldighed af antagelsen idet kausalitet er et forskningsområde i præstationsmålingsteorien, som ofte er blevet forsømt. Afhandlingens besked er at vi er i besiddelse af begrænset empirisk evidens for eksistensen af sande universelle kausalitetssammenhænge, der kan vejlede praksis gennem om generiske strategiske handlinger, der kan sikre succesfulde organisatoriske præstationer og

dermed danne rammen for forskrifter for ledelseshandlinger. Afhandlingen består af fire videnskabelige artikler og deres bidrag er som følger:

Den første artikel analyserer publiceret positivistisk økonomistyringsforskning for empiriske beviser på eksistensen af kausalitet i moderne præstationsmålingslitteratur. Ved at analysere progressionen i positivistisk økonomistyringsforskning på afdækningen af konsistente kausale relationer mellem ikke-finansielle måltal og finansielle præstationer, skaber vi klarhed på validiteten af antagelsen om kausalitet i moderne præstationsmåling. Afslutningsvis finder vi, at det empiriske grundlag for at påstå, at der eksisterer kausale relationer mellem ikke-finansielle måltal og finansielle præstationer, er i bedste fald svagt, da der er en klar mangel på konsistens i de empiriske resultater på dette. Vi er derfor ikke i stand til at tilvejebringe den nødvendige empiriske evidens for at kunne retfærdiggøre transformationen af kausalitetsantagelsen fra et *brute* fact til et *stylized* fact.

Den anden artikel analyserer, om publikationspraksissen i positivistisk økonomistyrings forskning tillader, at fænomenet questionable research practices (QRPs) kan slå rod. Problemet med QRPs er en fordrejning af den hypotetisk-deduktive metode, til fordel for en forskers egne hypoteser med bivirkningen at sandsynligheden for at finde et falsk-positivt resultat er øget. Hvis publikationspraksissen for positivistisk økonomistyringsforskning (utilsigtet) motiverer QRPs, må vi forvente, at QRPs er til stede, da det er et fænomen som er fundet at være almindelig udbredt i naturvidenskaben såvel som socialvidenskaben. Afslutningsvis finder vi, at den nuværende publikations tradition inden for positivistisk management accounting forskning skaber det nødvendige rum der skal til for at QRPs kan tage form, og vi tager derfor forbehold for at ratioen af falsk-positive kan være væsentlig over det antagne niveau på 5 procent. På denne baggrund rejser vi spørgsmålet om, hvorvidt det er videnskabeligt forsvarligt at advokere for kausalitet mellem ikke-finansielle måltal og finansielle præstationer når den empiriske evidens ikke er konsistent og den evidens vi har, er formentlig inficeret af en højere ratio af falsk-positive end forventet.

Den tredje artikel studerer brugen af præstationsmåling i den danske offentlige sektor. Artiklen udforsker hvordan efficiens og effektivitetskriterierne relaterer sig til den mangelfulde implementering af præstationsmåling i den danske offentlige sektor. Artiklen finder at på trods af bestræbelserne på at udvikle et velfungerende og vellykket præstationsmålingssystem, så fejler det i at opfylde sit formål. Hvilket er at styre handlinger og aktiviteter i retning af opnåelsen af strategiske målsætninger. Præstationsmålingssystemet ender med ikke at være i stand til at skabe intern transparens og dermed kan efficiens og effektivitet ikke blive balanceret igennem management kontrol. Resultat er et præstationsmålingssystem der mest af alt er en administrativ byrde, der skaber en falsk følelse af sikkerhed i optimeringen af de knappe ressourcer.

Den fjerde artikel er et casestudie der studerer validiteten af PMS-værktøjet 'mål og resultatstyring', som er udviklet af Moderniseringsstyrelsen. Artiklen er et studie af hvordan

præstationsmodeller i sig selv kan indeholde mangler der skaber problemer med validitet og hvordan disse problemer udfolder sig i praksis. Artiklen imødekommer et behov for at skabe en bedre forståelse af hvorfor præstationsmåling, som del element i new public management, ser ud til kontinuerligt at fejle i den danske offentlige sektor. Vores analyse af 'mål og resultatstyring' viser, at det er dårligt skitseret med uoverensstemmelser i den konceptuelle struktur, der fører til et illusionsbaseret sprogspil. Disse problemer bliver ikke adresseret i den faktiske anvendelse af 'mål og resultatstyring' mellem Uddannelses- og forskningsministeriet og de danske universiteter. Et af de strategiske objektiver i præstationskontrakterne var at forbedre uddannelseskvaliteten gennem et skema af push kausalitet, men det var ikke muligt at fastslå, hvilken opfattelse af kvalitet universiteterne stræbte mod eller hvilken type kvalitet ministeriet ønskede dem at stræbe imod. Dette resulterede i en mangeartet målformulering, der virkede mere baseret på indtryk end rational begrundelse, hvilket resulterede i at kontrakterne var åbne for fortolkning. Vi har med andre ord ingen anelse om, hvorvidt universitetsuddannelserne er af stigende kvalitet eller høj kvalitet, hvis vi kun ser på måltallene i kontrakterne. Opsummerende fandt vi, at 'mål og resultatstyring' ikke lever op til de grundlæggende principper for at levere et koncept, der kan facilitere formålet med at skabe efficiente og effektive offentlige institutioner.

TABLE OF CONTENTS

1. Introduction

1.1 Theoretical positioning in performance measurement theory

By nature, the concept of performance measurement is diverse. Researchers with different functional backgrounds such as accounting, operations management, marketing, finance, economics, psychology and sociology are all actively working with performance measurement (Neely, 2007). It provides a fascinating richness, but also results in performance measurement being an elusive concept, as its very definition and understanding depends on the functional background. This dissertation, however, perceives and explores performance measurement from the perspective of accounting.

Classical accounting theory defines performance measurement as the evaluation of organisational performance in economic terms (Ijiri, 1975), which implies that the evaluation was traditionally conducted with financial measures (Eccles, 1991; Otley, 2007). This meant that the evaluation of performance was reliant on the institutional formula of accounting for which financial measures are defined and derived (Lueg & Nørreklit, 2013). For example, the profitability of an organisation or managerial action is only indirectly observable through calculations based on the logic of accounting formulas (Lueg & Nørreklit, 2013; Malina, Nørreklit, & Selto, 2007). This rendered performance measurement to be a tool for providing information on the evaluation of strategies and actions in terms of financial performance, in other words, performance measurement was originally meant to be a tool to evaluate historical financial performance (Eccles, 1991; Lueg & Nørreklit, 2013). Traditional performance measurement systems (PMS) were therefore consisting of multiple financial measures, for example the '*pyramid of ratios*', which is a decomposition of the Return on Investment (ROI) measure (Otley, 2007). Other noteworthy financial measures that could be mentioned in this relation are for example Economic Value Added (EVA) and Net Present Value (NPV) (Lueg & Nørreklit, 2013).

However, during the 1980s a dissatisfaction with financial measures was spreading and it resulted in the highly influential publication by Johnson and Kaplan (1989) with the title *'Relevance Lost: The Rise and Fall of Management Accounting'* criticising the historic nature of financial measures and arguing for non-financial measures to have a more significant influence in performance measurement. The construction of financial measures is based on institutional accounting formulas, which ensured that they reveal a great deal about the past actions of an organisation but nothing about its future. This is due to financial measures not emphasising any element that will lead to good or poor *future* financial results (H. Nørreklit, 2000). Financial measures are simply unable to encapsulate any uncompleted chains of actions, which extend beyond the time of measurement.

This created a desire for performance measurement to provide more future-oriented accounting information, thereby overcoming the reactive and historical nature of financial performance measurement. And, as a result, a great deal of attention has been turned towards

the development and use of non-financial measures of performance that could be used to both motivate and report on the performance of an organisation (Eccles, 1991; Neely, 1999; Otley, 2007). However, non-financial measures are actually not a new phenomenon. General Electric, for example, combined the use of financial and non-financial measurement as part of identifying their key corporate performance indicators in the 1950s (Eccles, 1991; H. Nørreklit, 2000). Similarly, a number of researchers have also earlier pointed at the relevance of non-financial measures (Anthony, Dearden, & Bedford, 1984; Argyris, 1977; Hopwood, 1973), however, at this point in time, non-financial measurement was generally characterised by being loosely coupled local systems guided by local needs and with no integration of the strategic objectives of the organisation (Mouritsen, Høholdt, & Jørgensen, 1996). The loosely nature of non-financial measures, in comparison with financial measures, was a result of such measures not being dependent on the logic of accounting formulas; there were no rules for their formulation.

All of this changed with the publication by Johnson and Kaplan (1989) and the conceptualisation of PMSs that linked non-financial measures to the strategy of an organisation (Cross, Lynch, & McNair, 1990; Grady, 1991; Kaplan & Norton, 1992). This resulted in the perception that PMSs including a non-financial measurement where considered to be more future oriented than traditional PMSs. However, the inclusion of non-financial measures lacked a compelling argument for why they could be considered leading indicators and drivers of future financial performance. A captivating argument came with the reconceptualization of the Balanced Scorecard (BSC) in 1996, which integrated financial and non-financial performance measures, by linking outcome measures and performance drivers in cause-and-effect relationships, which became a baseline assumption of contemporary performance measurement (Kaplan & Norton, 1996). In consequence, empirical postulates of generic strategic actions that were to drive successful business performance were now embedded in PMSs, thus making specific prescriptions for managerial actions that would necessarily lead to future success (Lueg & Nørreklit, 2013; H. Nørreklit, 2000).

When claiming that non-financial measures are leading indicators of future financial performance, research had overcome the dissatisfaction with the historical nature of performance measurement. And as a result, contemporary performance measurement[1] comprise the use of both financial and non-financial performance measures linked to the business strategy of an organisation (Franco-Santos, Lucianetti, & Bourne, 2012). Examples of contemporary PMS could be the BSC (Kaplan & Norton, 1996, 2001), the multi-criteria key performance indicators (Cheng, Luckett, & Mahama, 2007; Hall, 2008) and the performance prism (Neely, Adams, &

[1] It is important to stress that in the literature, the term 'contemporary performance measurement' is used interchangeably with other phrases such as 'integrated performance measurement' (Bititci, Carrie, & McDevitt, 1997), 'comprehensive performance measurement' (Hall, 2008), 'strategic performance measurement' (Ittner, Larcker, & Randall, 2003), or 'business performance measurement' (McAdam & Bailie, 2002).

Crowe, 2001; Neely, Adams, & Kennerley, 2002). The adoption of such systems has been on the up rise since their theoretical inception (Bourne, Neely, Mills, & Platts, 2003; Rigby & Bilodeau, 2009, 2015) and they exist in many variations (Speckbacher, Bischof, & Pfeiffer, 2003).

The modern understanding of performance measurement has become an ongoing and evolutionary process (Eccles, 1991). However, the evolution of arguing that future financial performance is reliant on the embedment of generic strategic actions and prescriptions for managerial actions developed on the notion of causality between non-financial measures and financial performance was unquestionably a controversial claim (Ittner & Larcker, 2003; Kaplan & Norton, 1996, 2001). Initially, the idea of causality received critique (H. Nørreklit, 2000, 2003), but it managed to remain persistent in both theory and practice (Franco-Santos et al., 2012; Hoque, 2014; Lueg & Nørreklit, 2013; Micheli & Mari, 2014). It was the ambition that organisations should develop causal performance models with explicit cause-and-effect relations between non-financial and financial measures (Ittner & Larcker, 2003). A common example of such a cause-and-effect relationship is the claim that customer satisfaction and product quality are *certain* drivers of future financial performance (Lueg & Nørreklit, 2013). By building performance measurement models on cause-and-effect relationships, it would debatably equip an organisation with a tool to *a priori* know what drives the financial performance and hence renders the PMS proactive of the future (De Haas & Kleingeld, 1999). The importance of non-financial measures as measures of future organisational performance therefore rests on the assumption of empirical verifiable causal relationships between non-financial measures and financial performance. On the other hand, if non-financial measures were found to be unrelated to organisational performance, it would instead reduce non-financial measures from being performance measures to just measures (Ijiri, 1975) and thereby question their relevance in PMSs.

The development provided the ability for PMSs to hold different and new roles in organisations (Otley, 2007). First, they could provide a tool for financial management. Second, they could provide an objective for overall organisational performance. Third, they could provide a means for control and motivation. The sole purpose of performance measurement was therefore not economic anymore, a reflection also made by Ijiri (1975). Unquestionable, there is an overlap between these different roles, but if managers and researchers do not recognise these different roles, the result can be significant confusion, especially when a PMS designed to fulfil one role is used for another.

The dissertation is therefore positioned in the midst of the discussion of causality between non-financial measures and financial performance. The dissertation is built around an investigation of causality from four different aspects, reflected in the four papers. In addition, we perceive performance measurement from the second role, namely the provision of an objective for overall organisational performance.

Having briefly outlined the theoretical positioning of the paper and the development from traditional financial performance measurement to contemporary performance measurement, we will use the next section to highlight some aspects of performance measurement theory, which are still underdeveloped and hence motivate the research aim of the dissertation. Subsequently, we will provide a philosophical underpinning of performance measurement theory in accounting and lastly present the conceptual structure of the four papers in the dissertation.

1.2 Motivation and identified issues

Performance measurement has by accounting researchers been investigated from an array of research methods, such as case study research (Malina et al., 2007), survey research (Hoque, 2005; Ittner et al., 2003), quasi-experimental research (Banker, Potter, & Srinivasan, 2000; Davis & Albright, 2004) and experimental research (Lipe & Salterio, 2000; Tayler, 2010).

The studies have focused on different levels of analysis, ranging from how PMSs affect the behaviour and performance of individuals (Hall, 2011) to investigating the effects of PMSs when taking into consideration the aspects of design, implementation and use (Speckbacher et al., 2003). From a research point of view, we therefore possess broad knowledge about why organisations adopt contemporary PMSs (Chenhall & Langfield-Smith, 1998; Henri, 2006; Hoque, 2014; Hoque & James, 2000), but we are less knowledgeable about the consequences of implementing and operationalising these systems (Lee & Yang, 2011; Micheli & Mari, 2014). Franco-Santos et al. (2012) and Hoque (2014) therefore for more research on the actual consequences of contemporary PMS implementations and operationalisations.

We follow this suggestion and approach it by investigating the ideal of using causality in performance measurement and its implications for practice, as the relevance of non-financial measurement in performance measurement is arguably dependent on its cause-and-effect relationship with financial performance. Which means that contemporary PMSs are empirically defined through causality instead of being solely dependent on accounting calculations.

In terms of actual implementations and operationalisations of contemporary performance measurement, we investigate this by using the public sector as a case. Public sector organisations are unable to build PMSs around financial measurement, as financial measures are useless in the measurement of outcome and effectiveness (Anthony & Govindarajan, 2003; Anthony & Young, 1999; Forbes, 1998; Modell, 2005; Ridley & Simon, 1938, 1943). The organisational effectiveness of public organisations is not reflected in conventional accounting statements as healthy finances contain no information on the utility of provided service, which means that the ultimate objective of public organisations cannot be the realisation of financial objectives (Kaplan & Norton, 2001). However, the conceptualisation of cause-and-effect between non-financial measures and organisational performance was not limited to financial performance (Kaplan & Norton, 2001). The argument was that non-financial measures, in general, were performance drivers of outcome, which according to Kaplan and Norton (2001) rendered the

potential for contemporary PMSs even greater in the public sector than private sector. Because public organisations could use PMSs, such as the BSC, to conceptualise what output and outcome the organisation intends to achieve, in other words the strategic objective, and then derive the non-financial measures that are presumed to be causally linked to these outcomes. Performance measurement in the public sector therefore presents itself as a special case, where the use of financial measures is problematic, which renders the causal relations between non-financial measures and outcome even more critical for its success.

It is therefore of particular interest to observe and analyse how the public sector approaches performance measurement also considering the strong critique of this component of NPM (Hood, 1991, 1995; Hyndman & Lapsley, 2016) in practices (Hood & Dixon, 2015a, 2015b; Kaspersen & Nørgaard, 2015; Møller, Iversen, & Andersen, 2016).

Within the following paragraphs, we provide a brief presentation of the evolution in the philosophy of performance measurement in the management accounting domain. This is needed to understand why the change to include non-financial measures in causal-based performance measurement is a drastic and important change in the theoretical foundation for accounting-based performance measurement and management.

1.3 Theory of performance measurement in accounting

For accounting philosophy, it was a radical shift to go from treating financial figures as the foundation for PMSs to treat them as one among a broader set of measures (Eccles, 1991). It signalled a paradigmatic change in performance measurement no longer being a calculative practice founded on accounting formulas, but instead dependent on empirical cause-and-effect relations and non-financial measures for which there existed no rules of formulation.

In the next sections, we first introduce the accounting measurement philosophy behind financial measures in order to understand how significant and influential a change this was. To do so, we draw upon the work of Yuji Ijiri and his *'Theory on Accounting Measurement'* from 1975 (Ijiri, 1975). We then draw on several sources to develop an understanding of the development and influence of non-financial measures and causality on performance measurement from a management accounting perspective.

1.3.1 Financial measurement

Performance measurement is not limited to the measurement of economic or financial goals, as it could just as well be concerned with social or engineering goals. However, the common ground is that the measurement of performance occurs with respect to set goals irrespective of whether they are of a financial, social or engineering nature and that they are not always cooperative; in fact, it is rather rare to be in a situation where the interest of the agent completely coincides with the interest of the principal.

According to Ijiri (1975), this has several implications for financial performance measurement. First, performance measures, if disclosed indiscriminately, may infringe upon the agent's privacy, as disclosure to external parties could hurt the competitive advantage, while disclosure to internal superiors may lead to over-control. Performance measures therefore typically represent sensitive information on internal processes. More broadly, performance measurement is the evaluation of performance of an organisational unit or corporate unit. The most typical financial performance measures encountered is of course income measurement in terms of income or profits[2]. At first glance, income or profit measurement are quite an ambiguous concept. However, its underlying logic is operationally measurable, as we are able to highlight the achievement of economic goals, just as we are able to do so within sports. Without being able to operationalise income measurement, capitalism or business would be just as inconceivable as the Olympic Games without sports records (Ijiri, 1975). The operationalisation of income measurement is actually one of the most important and fundamental contributions from the accounting practice, and, more generally, it means that any performance measure must be operationally measurable.

Performance measures have several stakeholders. An organisation would use them for decision-making purposes and setting future goals, but they are also to be communicated to stakeholders outside of the organisation, such as shareholders, creditors, governmental agencies and the public. This illustrates the relationship between the organisation and the recipient of the performance measures. To protect stakeholders in the use of performance measures, the measurement used must be highly standardised and verifiable, so that there can be no dispute over a performance measure generated by an accounting system. In addition, those who are responsible for the performance measurement process must be clearly delineated, so that measures are produced and verified in a responsible manner; an approach that protects those who are responsible for the process for any indiscriminate accusations.

These implications illustrate that there is likely to be more pressure to bias performance measures than other more neutral measures and that performance measures should be carefully constructed in order to protect them from such biases. In reality, this renders it impossible and unwise to discuss performance measurement without understanding the pressures that may be exerted by all stakeholders of the measure. The discussion also points towards the need for developing criteria for constructing performance measures that are unequivocal and unambiguous (Ijiri, 1975).

To mitigate the possibility of disagreeing over performance measures, Ijiri (1975) develops the concept of hardness to express the potential of disputing over a performance measure. A 'hard' measure is defined as a measure constructed in a way for which there can be

[2] Other types of measures could again be ROI, EVA, ROCE, NPV etc.

little disagreement, while a 'soft' measure is defined as a measure that can easily be translated in one direction or the other. For example, asset turnover is a hard measure, while customer satisfaction is a soft measure. To develop a hard measure, three ingredients are crucial. First, the measurement process must be founded on facts, if measures are based upon fictions, hypotheses, opinions or unverifiable facts it invites disagreement in the translation of performance. Second, the construction of the measure must be well-specified to enable all parties to judge unambiguously, which measurement rules for transforming facts into figures are justifiable and which are not. Third, the number of justifiable rules for calculating the measure must be restricted. If two legitimate rules produce drastically different numbers from the same set of facts, the numbers produced cannot be considered hard, as there is room for disputes. This does not imply that there cannot be alternative rules. However, it is necessary to specify the conditions under which each alternative is applicable so that under a given situation only one alternative can be considered legitimate.

The criterion of hardness is a special property of performance measurement, but performance measurement shares other common properties to measurement in general. In particular, the criterion of 'identifiability' is of relevance to performance measurement. As the primary purpose of measurement is to communicate and reflect the state of other things, figures that are produced as an output of measurement have no utility apart from their function to represent the state of the objective. Measures that are used to convey information about the state of something else are defined as surrogates, and things or phenomena that are represented by a surrogate are defined as principals. Therefore, in relation to performance measurement, a principal is what managers are primarily interested in, and surrogates interest us only as long as they provide information about the state of the principals.

An example of a complex surrogate could be a financial statement, since people are not only interested in financial statements because of their artistic beauty, but because they represent aspects of the financial state of an organisation. More simple surrogates could be EVA, ROI, ROA etc. They all represent some aspect of the financial state of an organisation. A last property related to identifiability, is the fundamental requirement that the stakeholders can decode the surrogate, so that the state of the principal can be reliably unfolded. Surrogates that cannot be decoded: *"is like food that is not edible, a clock that does not keep time or a car that does not run. These items may be useful in some respect but certainly not for their intended purposes"* (Ijiri, 1975, p. 41). Performance measures that are produced as an output of measurement therefore contain no utility in themselves apart from their ability to represent the state of other things, principals.

Financial measures are developed around this philosophical underpinning, which is why it signalled a significant paradigmatic change for performance measurement theory, when non-financial measures were to be considered as their equal in PMSs. The unrestricted formulation

of non-financial measures is unquestionable in clear contrast to the accounting philosophy behind the construction of financial measures, which in the end renders non-financial measures to be perceived as typically 'soft measures', and that invokes a question of reliability. This is also why researchers over time have raised caution against an uncritical use of non-financial measures in terms of risking to formulate *invalid* measures, which are measures that are unable to capture what they were supposed to do, in other words the *principal* (Ittner & Larcker, 2003). Non-financial measures are therefore in risk of not meeting the criteria of identifiability, which questions whether it is possible to decode the surrogate so that the state of the principle can be reliably unfolded.

In the next section, we dig deeper into the formulation of non-financial measures and look at how the lack of criteria for the formulation of non-financial measures has a potential to erode the assumption of causality between these measures and financial performance.

1.3.2 Non-financial measurement

With the revolution of non-financial measurement came the addition of elements such as customer satisfaction, product quality, market share and human resources as a part of measuring organisational performance (Eccles, 1991). It meant, on a macro level, that these measures became surrogates of *future* organisational performance on an equal footing with financial measures, which were surrogates of *current* organisational performance. This implied that non-financial measures were considered to be leading indicators of future financial measures, and that they became the answer to *"what measures truly predict long-term financial success in our business?"* (Eccles, 1991, p. 131).

Especially, two different streams of non-financial measurement gained attraction during the 1980s and 1990s. The first was a quality movement that gained momentum in 1980s, later termed Total Quality Management (TQM), and it perceived quality as a strategic weapon in the battle of competitiveness (Eccles, 1991). A result of this movement was that manufacturing companies submitted significant resources to develop and improve measures such as defect rates, response time, delivery commitment and so on. At that time, researchers and practitioners saw this movement as representing the most positive step taken in broadening the basis of business performance measurement. The second movement that gained momentum was the measurement of customer satisfaction, which occurred during the 1990s (Anderson, Fornell, & Lehmann, 1994; Eccles, 1991; Lueg & Nørreklit, 2013). The interest in customer satisfaction was triggered by strategies emphasising customer service, which was the same for the TQM movement, and it is claimed that as competition stiffens, TQM strategies will evolve into strategies based on customer service. In terms of measurement, this means that only the quality measures that are related to customer satisfaction will be continuously measured, while new customer satisfaction measures will arise in likes of customer retention rates, market share and perceived value of goods and services.

What is common for both non-financial measurement movements are a question of hardness and identifiability. In terms of hardness, it can be doubtful whether either quality or customer satisfaction measures can be constructed with such a reliability that they can be considered hard. The construction of quality measures would be based on facts, however, which measures to consider as surrogates for quality (i.e. the principal) are questionable. While, surrogates for customer satisfaction are typically both constructed on facts and opinions, the construction of the measures is not well specified, and there is not restriction on the justifiable rules for calculation of the measure. In the end, non-financial measures are often measuring something very complex that is difficult to define and measure in its properties. This is also why non-financial measurement in practice is often reduced to measuring the simplest and most easily measurable aspect of an activity or process being conducted e.g. customer satisfaction = retention rate; academic standing = impact factor of publications; employees' satisfaction = number of complaints expressed by employees (Micheli & Mari, 2014).

The issues with hardness and identifiability are therefore due to the problem of interpreting and evaluating the principle for example as observed with quality and customer satisfaction. However, this also renders the surrogate to become an expert, legitimate or authentic in interpretation and evaluation of the principle. This is represented by performance measurement in itself demonstrating this property, as it becomes the main concern of every interested party instead of the true state represented by the measurement (Ijiri, 1975). It also illustrates that the introduction of non-financial measurement resulted in a broader focus on objectives and measures (Eccles, 1991; Ittner & Larcker, 2003; Otley, 2007). When perceiving non-financial measurement from the logic behind financial measurement, we see that the particular formulation of non-financial measures is of the utmost importance, as in terms of performance measures there is likely to be a strong bias influencing their construction (Ijiri, 1975; Ittner & Larcker, 2003), which in combination with a lack of criteria provides reason for concern.

The importance of non-financial measures in performance measurement theory rests on the notion of 'universal relationship'/'cause-and-effect'. It implies that the strategic action X precedes profitability Y in time and that the observation of the strategic action necessarily implies the subsequent observation of the incident of profitability Y. This follows from Hume's argument of 'constant conjunction' in his definition of causality, which is that Y always follows X (Hume, 1975). Such connexions or assertions can only be proven through empirical observation. Hence, the two events strategic action X and profitability Y are therefore logical independent, which means that we cannot deduce from strategic action to profitability by logical reasoning (Hume, 1975; Lueg & Nørreklit, 2013). An example of a claimed causality in performance measurement is found in the balanced scorecard, which assumes that a generic set of strategies drives the causal path between the performance measures: *"organizational learning and growth → efficient internal business process → customer satisfaction and loyalty → financial success"* (Kaplan & Norton,

1996, p. 31). It is therefore no longer a calculation based on accounting formula that decides whether an investment in product quality, customer satisfaction, market share or human resources is profitable. Instead, it is a causal expectation that such investments are *always* profitable when the causal relationship has been empirically verified. The success of such generic strategies is therefore dependent on the ability to empirically identify and verify them and that they remain consistent through time and space. The persistence in time and space is, however, questionable, as there are examples of such fallacy.

Ittner and Larcker (1998) for example found that fewer than 55 percent of vice presidents of organisations could directly relate their quality measures to operational, productivity or revenue improvements, while only 27 percent could relate them to accounting returns, and no more than 12 percent could relate them to stock returns. Another example, a European producer of drive systems was the quality leader in their industry at the expense of production cycle time, which was three times higher than their competitors' (Kaiser & Young, 2009). Despite their marketing efforts, top managers concluded that customers were neither able to understand nor willing to pay a premium for the superb quality, they therefore cut quality, with no expense to customer satisfaction. This freed up working capital at a level of 5 percent of their annual revenue.

This is just two examples that contradict the generic assumption that quality through space and time is a leading indicator of future financial performance. The contradiction can also be explained from a microeconomic standpoint for which accounting theory is built. In microeconomic theory, there will be an optimum, for which it no longer makes sense to improve quality or customer satisfaction. This point occurs when the incremental increase in marginal utility (or willingness to pay) becomes smaller than the incremental increase in marginal costs. At this point, it is no longer profitable to increase quality or customer satisfaction.

Through a couple of empirical examples and microeconomic theory we have demonstrated that the assumption of causality is not set in stone. In consequence, the notion of empirical postulates of generic actions that are certain to drive successful business performance and hence make specific prescriptions for managerial actions that would lead to future success can be questioned.

Despite claiming causality, the performance measurement literature has largely avoided to define and discuss what was meant by it. In the next section, we therefore provide a short philosophical underpinning of what causality implies and if there are other relationships which could better explain the relationship between non-financial measures and financial performance.

1.3.3 A philosophical underpinning of causality

Oxfords dictionary define causality as *'the relationship between cause and effect'* and the *'principle that everything has a cause'*, while Merriam-Webster define it as *'the relation between a cause and its effect or between regularly correlated events or phenomena'*. However, to define and understand causality and what lies behind it, we draw upon the work of David Hume in *'An*

Enquiry concerning the Human Understanding' (1748)[3], as it is his criteria for causality, which are usually assumed within theory of science (Edwards, 1972; Simon, 1970; Slife & Williams, 1995) and also within theory of statistics in social sciences (Angrist & Pischke, 2014).

According to Hume (1975, p. 74), causality cannot be discovered without experience: *"it is impossible for us, by any sagacity or penetration, to discover, or even conjecture, without experience, what event will result from it, or to carry our foresight beyond that object which is immediately present to the memory or senses"*. This means that a causal relationship between X and Y cannot be logically reasoned, it can only be empirically observed and substantiated through experiences. Hume continues and argues that to claim a universal law or meta-law, a causation, one would have to rely on an endless reproduction of the same connexion of events regardless of time, place and context: *"when one particular species of event has always, in all instances, been conjoined with another, we make no longer any scruple of foretelling one upon the appearance of the other, and of employing that reasoning, which can alone assure us of any matter of fact or existence. We then call the one object, Cause; the other, Effect. We suppose that there is some connexion between them; some power in the other, by which it infallibly produces the other, and operates with the greatest certainty and strongest necessity"* (Hume, 1975, pp. 74-75). In this quote, Hume outlines the foundation for his definition, which is twofold by consisting of an internal and external impression. The first, *"A cause is an object followed by another, where all objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second had never existed"* and, the second, *"A cause is an object followed by another, and whose appearance always conveys the thought to the other"* (Hume, 1975, pp. 76-77).

What we can learn from his definition of causation is that if we attain knowledge of cause-and-effect, we would have *perfect* knowledge of their connexion, and this would allow one to predict and control future events by their causes (Hume, 1975). To achieve 'perfect' knowledge, is *the* purpose of positivistic management accounting research (Ittner, 2014; Lachmann, Trapp, & Trapp, 2017; Luft & Shields, 2014), as it is the ambition to predict the future and thereby control future events by their causes. In this sense, management accounting researchers have adopted a scientific approach in trying to uncover patterns and laws, while removing all notions of human intentionality with a firm belief in causal determinism for explaining organisational performance[4] (Micheli & Mari, 2014). The law of demand is an example of a law claimed to be

[3] The reason why the conceptual clarification on causality does not draw upon David Hume's *'A Treatise of Human Nature'* (1738) is the fact that he himself argued that he had been *'guilty of a very usual indiscretion in going to the press to early',* which was why the Treatise *'fell dead-born from the press without reaching such distinction as even to excite a murmur among the zealots' (Hume, 1975, pp. viii-ix).* This dissertation therefore draws upon the understanding of causality as outlined in *'An Enquiry concerning the Human Understanding'* (1948).

[4] Friedrich Von Hayek commented on this issue in his Nobel Memorial Lecture. He stated that economics, as well as other social sciences, is subject to the 'physics envy', which has lead researchers to draw

universal generalisable, which asserts that by reducing the price of a commodity, it will always increase its demand (Ryan, Scapens, & Theobald, 1992). If we translate this to performance measurement theory, then through a cause-and-effect relationship between a non-financial measure and a financial measure, we could predict and control future financial performance by controlling for the causes e.g. when controlling for customer satisfaction, we should be able to predict and control future financial performance. This is the ultimate ambition and objective of contemporary PMS.

However, causality is not the only type of relation to be considered, and in the following we will explain three other types of relationships that are of relevance, e.g. logic relations, finality relations, and the concept of 'construct causality'.

Logic relations are typically represented through accounting models and NPV calculations within performance measurement serving the purpose of creating financial rationality (H. Nørreklit, 2000). For logic relations, it is accounting formula that defines them and not empirical observations of company structures, this implies that a relationship between X and Y can only be logically reasoned, in other words, they can be logically proved or rejected. Accounting is a system of logical relationships based on our perception that to drive a business is about making profits. In the sense, that when revenue increases more than costs, it means that profits are increased. The correctness of a financial result is therefore not proven through empirical observations, but instead calculated or reached through the use of accounting calculus.

Correlations that do not exhibit the characteristics of causation can in some cases be defined as finality relations, which are something that occurs when human actions, wishes and views are related to each other. A finality relation is defined as (I) *"a person believes a given action to be a means – the best means – to an end"* and (II) *"the end and this view actually cause the action"* (Føllesdal, Walle, & Elster, 1997, pp. 170-171). A finality relationship is therefore a reciprocal relationship, which implies an involvement between ends and means. The relation is therefore not external given, but due to human volition. For example, a satisfactory financial result may be optioned by first supplying a good product for a low price, rendering customers very satisfied and achieving a market share and an image, and then later raising the prices while customer satisfaction is reduced. This example implies that there are many possible means to reach and end, and each means may have numerous other effects (Arbnor & Bjerke, 1994). A finality relation does therefore not imply the existence of a general law, so claiming finality is more unambiguous than invoking causation. In contrast to causation, a finality relation is dependent on time, space and context and can therefore not be universally generalised.

A last type of relationship stems from a much more recent interpretation of practice relations and is termed *'construct causality'*, which is a product of pragmatic constructivism (H.

inappropriate conclusions and forcefully adopt methodologies and methods from physical sciences. Van Hayek refers to this issue as the 'scientistic error' (Von Hayek, 1989)

Nørreklit, Nørreklit, & Mitchell, 2010). The argument of pragmatic constructivism and construct causality is that business success is not dependent on adapting to any meta-laws but instead a result of organisational actors' ability to construct a joint set of successful relationships to the world in which they operate (H. Nørreklit, 2017; H. Nørreklit, Raffnsøe-Møller, & Mitchell, 2016). The argument is that a successful action is dependent on the organisational actors to establish a joint set of functioning activities, which produces a certain intended outcome. A PMS should therefore not be dependent on identifying universal cause-and-effect relations. Instead, it is something created in the local practices by the human actors through a system of operation generalisations in the establishment or construction of local causalities (H. Nørreklit, Nørreklit, & Mitchell, 2016; H. Nørreklit, Raffnsøe-Møller, et al., 2016). It is therefore the claim of pragmatic constructivism and construct causality that there exist no meta-laws for the relationship between non-financial measures and financial performance. Such a relationship would be forever changing depending on the organisational actors' ability to construct these relationships.

The nature of understanding a relationship between two phenomena is dependent on ontology and epistemology, as these concepts determine reality and how knowledge of reality can be acquired. The method of studying any such relationship is therefore also dependent on the ontology and epistemology and we therefore use the next section to explain the methodological perspective of the dissertation and each of the four papers.

1.4 Philosophy of science: A methodological perspective

A paradigm is a way of describing a worldview through a set of philosophical assumptions about: the *nature of social reality* (known as ontology – what do we believe about the nature of reality?), *ways of knowing* (known as epistemology – how do we know what we know?) and *ethics and value systems* (known as axiology – what do we believe is true?) (Patton, 2002). A given paradigm thus leads researchers to ask certain questions and then uses an appropriate approach to the scientific inquiry (known as methodology – how do we study the world?).

Burrell and Morgan conceptualised these dimensions onto an 'objective-subjective' continuum, where one end stresses the objective nature of *reality, knowledge and human behaviour*, while the other emphasises the subjective aspects.

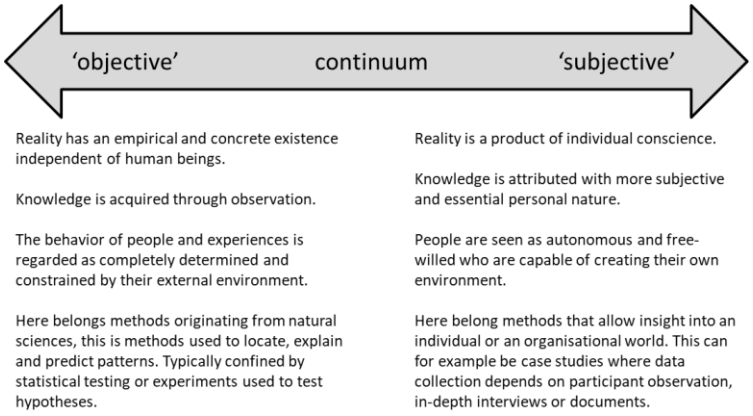


Fig. 1. The 'objective-subjective' continuum built on the Burrell and Morgan framework (Burrell & Morgan, 1979).

As such, *Ontology* is concerned with the nature of 'reality', where on the one hand, the social world and its structures can be regarded as having an empirical, concrete existence external to, independent of and prior to the perception of any human being. At the other extreme, reality is a product of individual consciousness, i.e. that the external social world consists of concepts and labels created by people to create a shared conception of its nature with others (Burrell & Morgan, 1979). In this sense, ontology is related to whether we believe there is one verifiable reality or if there exist multiple, socially constructed realities (Patton, 2002). *Epistemology* is concerned with the nature of knowledge, in what forms it can take and how it can be obtained and transmitted. Epistemology therefore asks the following questions: what are the sources of knowledge? How reliable are these sources? How do we know if something is true? What methods do we use to analyse a phenomenon? At the one extreme, knowledge can be acquired through observations and built piece by piece, while at the other extremity, knowledge is attributed to a more subjective and essentially personal nature. The Burrell and Morgan (1979) framework also included a discussion of *the human nature* referring to the relationship between a human being and their environment. At one end of the continuum, people's behaviour and experiences can be regarded as completely determined and constrained by their external environment or, on the other hand, people can be viewed as being potentially autonomous and free-willed who are capable of creating their own environment.

Together, these paradigmatic dimensions determine the assumptions and beliefs that frame a researcher's view of a research problem or phenomenon, how the researcher goes about investigating it, and the methods used to answer the research questions. If the social world is treated as the physical or natural world, then we must apply methods originating from natural sciences. These are methods that tend to be used to locate, explain and predict regularities and patterns, typically confined by statistical or experimental techniques used to test hypotheses and analyse collected data by standardised statistical instruments such as different types of regression analysis, structural equation modelling, experimental testing etc. Alternatively, if subjectivity in the form of experiences of individuals and the creation of a social world is acknowledged, then

methods that allow insight into an individual or organisational world are to be emphasised. This can for example be case studies where data collection depends on participant observation, in-depth interviews or documents.

Researchers tend to prescribe to a certain paradigm, thereby letting the paradigm define themselves as researchers (Weick, 1996). In this sense, you are either quantitative (e.g. logical positivist) or qualitative (e.g. social constructivist) resulting in a black or white perception of reality where the world is either ontological objective or subjective. This means that from an ontological objective perspective, phenomena only exist independent of human actors and the opposite if perceived from an ontological subjective perspective. Such a situation results in an inability for the world to hold both subjective constructions and objective phenomena and it also carries the unfortunate consequence that a researcher would approach all problems with the same method instead of letting the problem define the method needed. In the words of psychologist Maslow (1966, p. 15), *"if all you have is a hammer, everything looks like a nail"* meaning that if you 'are' a positivist, then the only hammer you possess is 'the scientific method' (statistical and experimental analysis), which renders all problems to be hypotheses that are to be tested.

However, we argue that one research approach does not fit all research questions and that reality or the world consists of multiple types of phenomena and constructions so we are limiting ourselves in understanding the world if we limit ourselves to one paradigm with one type of method. One should instead allow a problem to inform the researcher about which method is needed, as one paradigm might be more useful for answering one research question than another.

The world view of this dissertation is *pragmatic constructivism* (H. Nørreklit, 2017; H. Nørreklit et al., 2010; H. Nørreklit, Raffnsøe-Møller, et al., 2016; L. Nørreklit, Nørreklit, & Israelsen, 2006), as it can apprehend multiple methodologies due to pragmatic constructivism accepting that the world consists of phenomena ranging from the natural laws of physics to the social constructs of money or performance measurement models. This approach also allows the dissertation to vary in the type of research questions asked and the appropriate method applied. We acknowledge that reality consists of multiple types of phenomena such as *human beings, physical and biological phenomenon, social and physical constructs etc.* (see figure 2 for more details on what constitutes reality). The main purpose of pragmatic constructivism is to understand practice and therefore how to *'make things work'* and to search for a pragmatic truth – did it work?

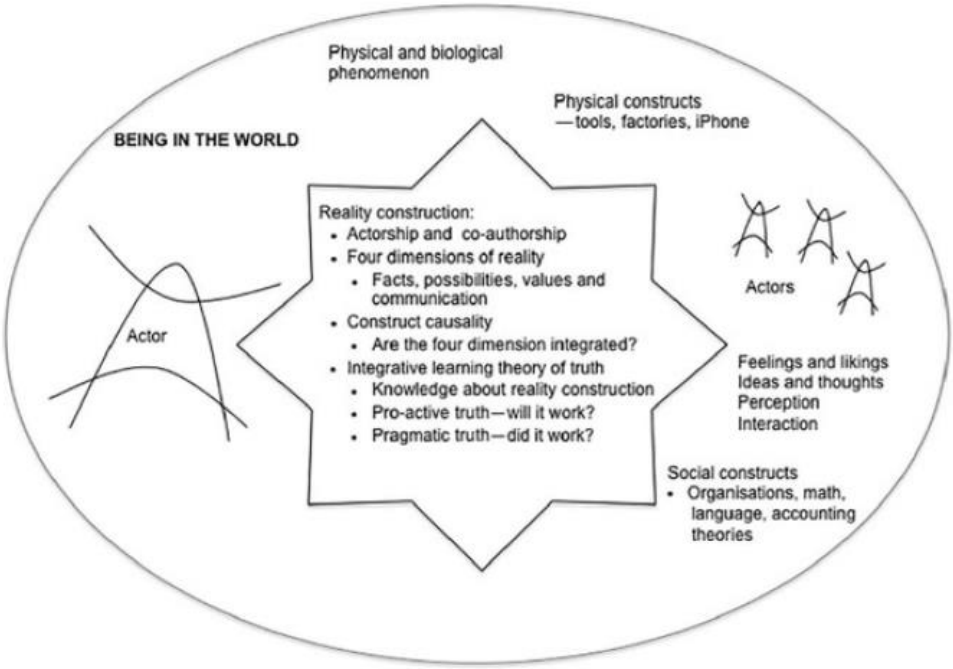


Fig. 2. Pragmatic constructivism's view on what constitutes reality i.e. the learning circle between proactive truth and pragmatic truth (H. Nørreklit, 2017)

However, before we can have pragmatic truth, we are left with a proactive truth, which is the presumed truth as we see it before we take action (L. Nørreklit, 2017). Proactive truth is the here-and-now perspective based on existing evidence, anticipation and adequate analysis, while the pragmatic truth is based on the future fulfilment of the expectations that the truth claim produce. Something is only pragmatic true if the expectations become fulfilled in future observations. It is therefore neither a forecast nor a prediction. We will explain it with an example from L. Nørreklit (2017, pp. 92-93). The claim that 'it is raining' is based on observations, which is a proactive truth. In addition, 'it is raining' makes us expect that we will get wet if we step out. This is a concern for the future and if it holds true, then the claim is pragmatically true. This translates into two interpretations of the truth of the same claim "*one that is historical and one that is oriented at the future and carries information that is relevant to the formation of intentional actions. Both of them are based on evidence – existing evidence and possible future evidence"* (L. Nørreklit, 2017, p. 93).

Overall, the dissertation is searching for the pragmatic truth of causality in contemporary performance measurement. Does the assumption hold? Can it be generalised to practice? To do so, we need a highly detailed and complex representation of knowledge about the assumption of causality. To address this, pragmatic constructivism offers an integrative learning theory of truth that is intended to enable us to theoretically point to problems of validity. The end result should be a learning theory of truth where the learning circle, i.e. the interplay between the proactive truth of whether the projection will hold true, and the pragmatic truth of whether it did hold true, forms the basis of the learning process (H. Nørreklit, 2017). To create a learning theory of truth,

one must therefore create an interplay between the conception of pragmatic truth and the more conventional pro-active truth, which is the main reason behind the division of the thesis into two sections.

To uncover the pragmatic truth of causality we have divided the dissertation into two sections. The first section is consisting of paper one and two, and analyses the progress of research in finding and validating specific causal relations. This part of the thesis is searching for the proactive truth of causality in performance measurement. These two papers are discussing the here-and-now perspective of existing evidence for causal claims: Specifically, we look for researchers' observations and consistency in the findings of causality as well as the reliability and validity of the methods currently employed to uncover any causations. The empirics of these two papers are published literature employing methods of either statistical or experimental nature. The choice of literature is consistent with the epistemological assumption of causality, as methods for studying causations originates from natural sciences, where they are used to locate, explain and predict patterns i.e. causations. Furthermore, we look for consistency, because to identify a correlation to be a true causation, it would require the correlation to be infinitely reproducible: *"even after one instance or experiment, where we have observed a particular event to follow upon another, we are not entitled to form a general rule, or foretell what will happen in like cases; it being justly esteemed an unpardonable temerity to judge of the whole course of nature from one single experiment, however accurate or certain"* (Hume, 1975). In the end, the first two papers address how far research is in substantiating the claim of causality and thereby moving it from the hazy zone of uncertain speculation to presumed certainty. In other words, the papers provide insight into the extent to which the research making pro-active truth statements on natural science causality has shown to be pragmatic true

The second section, consisting of the last two papers, investigates causality from the question 'what about practice?' This section is concerned about the pragmatic truth of causality statements in how practice approaches the notion of causality in performance measurement. This is done by analysing actual case implementations of PMS that employ non-financial measures. This approach is taken due to the argument of pragmatic constructivism that a successful organisational practice is not given by nature and cannot simply be managed and measured routinely from the vantage point of management. Instead, one would have to investigate, understand and theorise on functional practices as constructed through the activities of the human actors in their construction of reality. This implies that the empirics of the last two papers must reflect this construction of reality and we therefore chose to focus on observations from mainly documents and interviews. We search for information that is relevant to the formation of intentional actions. What is of interest here is the pragmatic truth of the PMS implementations; are the practitioners in the two cases able to construct PMS with valid causations between non-financial measures and the objective of the organisations?

Taken together, these two approaches provide the basis for learning about the validity of the assumption of causality to exist between non-financial measures and financial performance in performance measurement theory.

1.5 Structure of the dissertation

The dissertation consists of four papers that will be explained in more detail in the following.

Paper I. **CAUSALITY IN CONTEMPORARY PERFORMANCE MEASUREMENT: ARE CAUSAL QUESTIONS BEING ANSWERED?**	
Author(s)	Kristian Mohr Røge
Research Question(s)	How far has PMAR come in providing consistent empirical answers to causal questions(s) of non-financial measures being leading indicators of future financial performance?
Approach	Systematic review of normal science on causation in contemporary performance measurement literature

Paper one aims at systematically reviewing the consistency of empirical evidence on causality in contemporary performance measurement literature, as we intend to uncover the profoundness of the causality presumption. In doing so, we analyse the progress of positivistic management accounting research (PMAR) on providing exemplars (Kuhn, 1970) that can provide the assumption of causality with empirical content. The empirical material in the analysis is a carefully selected range of positivistic performance measurement studies from 1996 to 2014. If causality is found to be consistent in the analysed papers and particularly in the exemplars, we provide concrete evidence on the existence of cause-and-effect relations between non-financial measures and financial measures, so that the brute fact of causality in contemporary performance measurement can be translated as a stylized fact instead of a brute fact. However, findings of the paper evidence that the empirical ground for claiming causality between non-financial measures and financial performance is inconsistent and unconvincing.

The paper consequently calls for additional studies that could inform if the lack of consistency is due to methodological difficulties or a lack of appreciating corroborating studies in the publication environment. Thus, the next paper is aimed at investigating the publication culture of doing scientific inferences in PMAR.

Paper II. **IS THE VALIDITY OF POSITIVISTIC MANAGEMENT ACCOUNTING RESEARCH EXPOSED TO QUESTIONABLE RESEARCH PRACTICES?**	
Author(s)	Kristian Mohr Røge
Research question(s)	How susceptible are the publication practices of PMAR to the phenomenon of QRPs?
Approach	Meta-analysis of published PMAR from 2010-2015

The second paper takes point of departure in analysing published PMAR from 2010 to 2015 for indications of a publication practice that allows for QRPs to take root. The concern is that QRPs are found to be widespread within natural and social sciences and QRPs are argued to distort the hypothetico-deductive method in favour of a researchers own hypothesis with the side-effect of increasing the likelihood of experiencing a false-positive. This is problematic for PMAR, as null-hypothesis testing (NHST) is considered the *sine qua non* method of scientific inference in PMAR. If the publication practices of PMAR are unintentionally allowing for QRPs, we would expect QRPs to be present. It is unfortunately a phenomenon that contaminates the validity of the hypothetico-deductive method in making reliable causal claims, as it distorts the knowledge pool in an almost undetectable way. We therefore investigate the publication practices of PMAR and not per se for QRPs in itself.

In conclusion, we find that the current publication practice of PMAR provide space for QRPs to flourish, and we would therefore expect the ratio of false-positives in PMAR to be well above the assumed ratio of 5 percent. We question if the advocacy of causalities between non-financial measures and financial measures is sound considering that the empirical evidence is inconsistent and that the ratio of false-positives is expectedly higher than the conventional ratio of 5 percent.

The last two papers are case studies of contemporary performance measurement in the Danish public sector. We chose to conduct these studies within the public sector due to this sectors inability to use financial performance in appraising organisational performance. Performance measurement in the public sector is therefore reliant on the ability to formulate outcome measures that are linked to input, process and output measures, we investigate how this unfolds in practice and how causality interplay in this.

Paper III. **A STUDY ON THE CRITERIA OF INTERNAL TRANSPARENCY, EFFICIENCY AND EFFECTIVENESS IN MEASURING LOCAL GOVERNMENT PERFORMANCE**	
Author(s)	Kristian Mohr Røge and Niels Joseph Lennon
Research Question(s)	To what extent has the management of the municipality met the criteria of efficiency and effectiveness in their performance contracts between top management and organisational units, and what role does the attainment of the criteria of measurement play in ensuring a functioning PMS?
Approach	Case study
Case	A Danish municipality

For a long time, the public sector has been exposed to allegations of wastefulness and inefficiency. As a result, performance measurement has become an indispensable and universal element for modernising local government entities in achieving 'more for less'. Performance measurement is considered a tool enabling public managers to answer the following questions: (1) *'how adequate and effective is our service performance?',* and (2) *'how efficient are we in providing these services?'.* However, empirical results have evidenced that the implementation of PMS in the public sector is rarely a success, as it has not resulted in the expected improvements in performance, accountability, transparency and quality of services.

It has been argued that the inadequacy of performance measurement is due to an unresolved issue with formulating performance measures. This problem is attributed to the lack of a single, satisfactory, overall measure of performance comparable to the measurement of the financial performance of private organisations along with the intangible nature of public services. For public organisations, it is easy to measure how much work has been done – but not how well it was done, nor whether the particular work undertaken was appropriate to the desired end. These factors put even more pressure on the formulation and use of non-financial measures in measuring organisational performance.

This paper explores how the efficiency and efficiency criteria relate to the inadequacy of PMS implementations in local government entities. We argue that a measurement of these two criteria must be central for public sector PMSs to achieve the strategic objectives with an efficient resource consumption under financial constraints, as the efficiency and effectiveness criteria are what render the resource flow from costs, through outputs, to outcome transparent and manageable.

We find that notwithstanding the endeavours to develop a well-functioning and successful PMS, the analysis argues that the PMS fails in accomplishing its purpose of direction, actions and activities toward the achievement of strategic objectives. It ends up being a PMS that is unable to create internal transparency, and therefore efficiency and effectiveness cannot be balanced

through the activities of management control. The PMS becomes an administrative burden that provides top management with a false sense of security in the optimisation of scarce resources.

The study provides evidence for the argument that the linkage between non-financial measurement and outcome measures is something that is constructed in the local practices. Instead of being dependent on universal causations that are generalised from empirical postulates, found by scientific inferencing.

Paper IV. **THE ILLUSION OF 'OBJECTIVE AND RESULT-BASED MANAGEMENT': A UNIVERSAL NPM TOOL IN THE DANISH PUBLIC SECTOR**	
Author(s)	Kristian Mohr Røge, Nikolaj Kure, Hanne Nørreklit
Research Question(s)	(1) What characterises the conceptual qualities embedded in the model of objective and result-based management? (2) Given the pragmatic constructivist definition of validity as described above, do these conceptual qualities facilitate a valid model that may stimulate construct causality (3) How may we explain that a language game of illusion exists in the realms of performance management of the Danish Public sector?
Approach	Case study
Case	A combination of the contractual relationship between the Ministry of Higher Education and Science and Danish Universities and the steering logic promoted by the Agency of Modernisation.

In recent years the Agency of Modernization has accelerated the deployment of NPM tools by refining its standards and expanding its range of operation. A key component of this acceleration is the use of management accounting schemes that aim to visualise 'unused' resources and redirect them to more cost-efficient initiatives.

In this paper, we analyse one of these tools, namely 'objective and result-based management', which is a performance measurement and management system constructed as a contractual relationship between parties. Typically, between a ministry and a governmental agency, such as a hospital or university with the purpose of increasing the efficiency and effectiveness of public services. The paper aims at evaluating the validity of 'objective and result-based management'. It is an examination of how a performance model in itself may contain deficiencies that dispose for problems of validity and then how these validity issues unfold in practice. The study is divided into two sections. The first is an analysis of the theoretical underpinning of the model, as outlined in the material developed by the Agency of Modernization. The second section is a study of its implementation between the Ministry of Higher Education and Science and seven Danish universities. We address a need for a better understanding on why the implementation of such NPM tools appears to be continuously failing in the Danish public sector.

Our analysis evidences that 'objective and result-based management' is poorly outlined and with mismatches in its conceptual structure that leads to a language game of illusions. This

study also shows how the practice of founding performance measurement on causal schematics, which involves management prescriptions of right actions leading to desired results linking action to results failed when implemented in a complex local context such as a university setting. In addition, our analysis showed that the causal schematic outlined was obviously too vague, uncertain and general to guide actions that lead to the desired end. We therefore conclude, that the outlined performance management framework 'objective and result-based management' do not live up to the basic principles for providing concepts that can facilitate the purpose of creating effective public sector institutions.

The next four chapters of the dissertation will present each of the four articles and chapter six will present the overall conclusion, contribution and practical implications for the dissertation.

References

Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from Sweden. *The Journal of Marketing*, 53-66.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering'metrics: the path from cause to effect*. New Jersey: Princeton University Press.

Anthony, R. N., Dearden, J., & Bedford, N. M. (1984). *Management control systems*. IL: McGraw-Hill/Irwin.

Anthony, R. N., & Govindarajan, V. (2003). *Management control systems* (11. international edition ed.). Boston, Mass.: McGraw-Hill/Irwin.

Anthony, R. N., & Young, D. W. (1999). *Management control in nonprofit organizations* (Vol. 6): Irwin Homewood, IL.

Arbnor, I., & Bjerke, B. (1994). *Företagsekonomisk metodlära*: Studentlitteratur.

Argyris, C. (1977). Organizational learning and management information systems. *Accounting, Organizations and Society, 2*(2), 113-123.

Banker, R. D., Potter, G., & Srinivasan, D. (2000). An empirical investigation of an incentive plan that includes nonfinancial performance measures. *The Accounting Review, 75*(1), 65-92.

Bititci, U. S., Carrie, A. S., & McDevitt, L. (1997). Integrated performance measurement systems: a development guide. *International journal of operations & production management, 17*(5), 522-534.

Bourne, M., Neely, A., Mills, J., & Platts, K. (2003). Implementing performance measurement systems: a literature review. *International Journal of Business Performance Management, 5*(1), 1-24.

Burrell, G., & Morgan, G. (1979). *Sociological paradigms and organisational analysis* (Vol. 248): london: Heinemann.

Cheng, M. M., Luckett, P. F., & Mahama, H. (2007). Effect of perceived conflict among multiple performance goals and goal difficulty on task performance. *Accounting & Finance, 47*(2), 221-242.

Chenhall, R. H., & Langfield-Smith, K. (1998). The relationship between strategic priorities, management techniques and management accounting: an empirical investigation using a systems approach. *Accounting, Organizations and Society, 23*(3), 243-264.

Cross, K., Lynch, R., & McNair, C. (1990). Do Financial and Non-Financial Measures Have to Agree? *Management Accounting*(November), 28-39.

Davis, S., & Albright, T. (2004). An investigation of the effect of balanced scorecard implementation on financial performance. *Management Accounting Research, 15*(2), 135-153.

De Haas, M., & Kleingeld, A. (1999). Multilevel design of performance measurement systems: enhancing strategic dialogue throughout the organization. *Management Accounting Research, 10*(3), 233-261.

Eccles, R. (1991). The performance measurennent manifesto. *Harvard business review, 69*(1), 131-137.

Edwards, P. (1972). The encyclopaedia of philosophy (Vols 1–8). US: Macmillian Publishing Co., Inc. & The Free Press.

Forbes, D. P. (1998). Measuring the unmeasurable: Empirical studies of nonprofit organization effectiveness from 1977 to 1997. *Nonprofit and voluntary sector quarterly, 27*(2), 183-202.

Franco-Santos, M., Lucianetti, L., & Bourne, M. (2012). Contemporary performance measurement systems: A review of their consequences and a framework for research. *Management Accounting Research, 23*(2), 79-119.

Føllesdal, D., Walle, L., & Elster, J. (1997). Argumentasjonsteori. *Sprak Og*.

Grady, M. W. (1991). Performance measurement: implementing strategy. *Strategic Finance, 72*(12), 49.

Hall, M. (2008). The effect of comprehensive performance measurement systems on role clarity, psychological empowerment and managerial performance. *Accounting, Organizations and Society, 33*(2), 141-163.

Hall, M. (2011). Do comprehensive performance measurement systems help or hinder managers' mental model development? *Management Accounting Research, 22*(2), 68-83.

Henri, J.-F. (2006). Organizational culture and performance measurement systems. *Accounting, Organizations and Society, 31*(1), 77-103.

Hood, C. (1991). A public management for all seasons? *Public administration, 69*(1), 3-19.

Hood, C. (1995). The "New Public Management" in the 1980s: variations on a theme. *Accounting, Organizations and Society, 20*(2-3), 93-109.

Hood, C., & Dixon, R. (2015a). *A government that worked better and cost less? : evaluating three decades of reform and change in UK central government* (First edition ed.). Oxford: Oxford University Press.

Hood, C., & Dixon, R. (2015b). What we have to show for 30 years of new public management: Higher costs, more complaints. *Governance, 28*(3), 265-267.

Hopwood, A. G. (1973). *Accounting and human behavior*. New Jersey: Prentice-Hall.

Hoque, Z. (2005). Linking environmental uncertainty to non-financial performance measures and performance: a research note. *The British Accounting Review, 37*(4), 471-481.

Hoque, Z. (2014). 20 years of studies on the balanced scorecard: Trends, accomplishments, gaps and opportunities for future research. *The British Accounting Review, 46*(1), 33-59.

Hoque, Z., & James, W. (2000). Linking balanced scorecard measures to size and market factors: impact on organizational performance. *Journal of management accounting research, 12*(1), 1-17.

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals* (3. ed., repr. / with text rev. and notes by P.H. Nidditch ed.). Oxford: Clarendon.

Hyndman, N., & Lapsley, I. (2016). New Public Management: The Story Continues. *Financial Accountability & Management, 32*(4), 385-408.

Ijiri, Y. (1975). *Theory of accounting measurement*. Florida: American Accounting Association.

Ittner, C. D. (2014). Strengthening causal inferences in positivist field studies. *Accounting, organizations and society, 39*(7), 545-549.

Ittner, C. D., & Larcker, D. F. (1998). Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction. *Journal of accounting research, 36*, 1-35.

Ittner, C. D., & Larcker, D. F. (2003). Coming up short on nonfinancial performance measurement. *Harvard Business Review, 81*(11), 88-95.

Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, organizations and society, 28*(7), 715-741.

Johnson, H. T., & Kaplan, R. S. (1989). *Relevance lost : the rise and fall of management accounting* (2. print ed.). Boston, Mass.: Harvard Business School Press.

Kaiser, K., & Young, S. D. (2009). Need cash? Look inside your company. *Harvard business review, 87*(5), 64-71.

Kaplan, R. S., & Norton, D. P. (1992). The balanced scorecard--measures that drive performance. *70*(1), 71.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: translating strategy into action*: Harvard Business Press.

Kaplan, R. S., & Norton, D. P. (2001). *The strategy-focused organization: How balanced scorecard companies thrive in the new business environment*: Harvard Business Press.

Kaspersen, L. B., & Nørgaard, J. (2015). *Ledelseskrise i konkurrencestaten* (1. udgave ed.). Kbh.: Hans Reitzel.

Kuhn, T. S. (1970). The structure of scientific revolutions, International Encyclopedia of Unified Science, vol. 2, no. 2: Chicago: The University of Chicago Press.

Lachmann, M., Trapp, I., & Trapp, R. (2017). Diversity and validity in positivist management accounting research—A longitudinal perspective over four decades. *Management Accounting Research*.

Lee, C.-L., & Yang, H.-J. (2011). Organization structure, competition and performance measurement systems and their joint effects on performance. *Management Accounting Research, 22*(2), 84-104.

Lipe, M. G., & Salterio, S. E. (2000). The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review, 75*(3), 283-298.

Lueg, R., & Nørreklit, H. (2013). Performance measurement systems - beyond generic actions. In F. Mitchell, H. Nørreklit, & M. Jakobsen (Eds.), *The Routledge Companion to Cost Management* (pp. 342-359). Abingdon, Oxon: Routledge.

Luft, J., & Shields, M. D. (2014). Subjectivity in developing and validating causal explanations in positivist accounting research. *Accounting, organizations and society, 39*(7), 550-558.

Malina, M., Nørreklit, H., & Selto, F. (2007). Relations among measures, climate of control, and performance measurement models. *Contemporary Accounting Research, 24*(3), 935-982.

Maslow, A. H. (1966). *The Psychology of Science*. New York: Harper & ROw.

McAdam, R., & Bailie, B. (2002). Business performance measures and alignment impact on strategy: The role of business improvement models. *International journal of operations & production management, 22*(9), 972-996.

Micheli, P., & Mari, L. (2014). The theory and practice of performance measurement. *Management Accounting Research, 25*(2), 147-156.

Modell, S. (2005). Performance management in the public sector: past experiences, current practices and future challenges. *Australian Accounting Review, 15*(37), 56-66.

Mouritsen, J., Høholdt, J., & Jørgensen, A. A. V. (1996). De "nye" og de "gamle" ikke-finansielle nøgletal. *Økonomistyring & informatik, Årg. 11, nr. 6 (1995/1996)*, 387-409 ; [396 241-263].

Møller, M. Ø., Iversen, K., & Andersen, V. N. (2016). *Review af resultatbaseret styring*. Retrieved from København:

Neely, A. (1999). The performance measurement revolution: why now and what next? *International Journal of Operations & Production Management, 19*(2), 205-228.

Neely, A. (2007). *Business Performance Measurement: unifying theory and integrating practice* (2. ed.). Cambridge: Cambridge University Press.

Neely, A., Adams, C., & Crowe, P. (2001). The performance prism in practice. *Measuring business excellence, 5*(2), 6-13.

Neely, A., Adams, C., & Kennerley, M. (2002). *The performance prism: The scorecard for measuring and managing business success*: Prentice Hall Financial Times London.

Nørreklit, H. (2000). The balance on the balanced scorecard a critical analysis of some of its assumptions. *Management Accounting Research, 11*(1), 65-88.

Nørreklit, H. (2003). The balanced scorecard: what is the score? A rhetorical analysis of the balanced scorecard. *Accounting, organizations and society, 28*(6), 591-619.

Nørreklit, H. (2017). *A Philosophy of Management Accounting: A Pragmatic Constructivist Approach*. New York: Routledge.

Nørreklit, H., Nørreklit, L., & Mitchell, F. (2010). Towards a paradigmatic foundation for accounting practice. *Accounting, Auditing & Accountability Journal, 23*(6), 733-758.

Nørreklit, H., Nørreklit, L., & Mitchell, F. (2016). Understanding practice generalisation–opening the research/practice gap. *Qualitative Research in Accounting & Management, 13*(3), 278-302.

Nørreklit, H., Raffnsøe-Møller, M., & Mitchell, F. (2016). A pragmatic constructivist approach to accounting practice and research. *Qualitative Research in Accounting & Management, 13*(3), 266-277.

Nørreklit, L. (2017). Epistemology. In H. Nørreklit (Ed.), *A Philosophy of Management Accounting: A Pragmatic Constructivist Approach* (pp. 87-108). New York: Routledge.

Nørreklit, L., Nørreklit, H., & Israelsen, P. (2006). The validity of management control topoi: Towards constructivist pragmatism. *Management Accounting Research, 17*(1), 42-71.

Otley, D. (2007). Accounting performance measurement: a review of its purposes and practices. In A. Neely (Ed.), *Business Performance Measurement* (2. ed., pp. 11-35). Cambridge: Cambridge University Press.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3. ed. ed.). Newbury Park: Sage.

Ridley, C. E., & Simon, H. A. (1938). The criterion of efficiency. *The Annals of the American Academy of Political and Social Science, 199*(1), 20-25.

Ridley, C. E., & Simon, H. A. (1943). *Measuring municipal activities: A survey of suggested criteria for appraising administration*: The International City Managers' Association.

Rigby, D., & Bilodeau, B. (2009). *Management tools and trends 2009*. Retrieved from US:

Rigby, D., & Bilodeau, B. (2015). *Management tools & trends 2015*. Retrieved from US:

Ryan, B., Scapens, R. W., & Theobald, M. (1992). *Research method and methodology in finance and accounting*. London: Academic Press.

Simon, J. L. (1970). The Concept of Causality in Economics. *Kyklos, 23*(2), 226-254. doi:10.1111/j.1467-6435.1970.tb02556.x

Slife, B. D., & Williams, R. N. (1995). *What's behind the research?: Discovering hidden assumptions in the behavioral sciences*. London: Sage.

Speckbacher, G., Bischof, J., & Pfeiffer, T. (2003). A descriptive analysis on the implementation of balanced scorecards in German-speaking countries. *Management Accounting Research, 14*(4), 361-388.

Tayler, W. B. (2010). The balanced scorecard as a strategy-evaluation tool: the effects of implementation involvement and a causal-chain focus. *The Accounting Review, 85*(3), 1095-1117.

Von Hayek, F. A. (1989). The pretence of knowledge. *The American Economic Review, 79*(6), 3-7.

Weick, K. E. (1996). Drop Your Tools: An Allegory for Organizational Studies. *Administrative Science Quarterly, 41*(2), 301-313.

Chapter 2

CAUSALITY IN CONTEMPORARY PERFORMANCE MEASUREMENT: ARE CAUSAL QUESTIONS BEING ANSWERED?

Author: Kristian Mohr Røge

Abstract Contemporary performance measurement comprise the use of financial as well as non-financial performance measures linked to the organisation's strategy and is conceptualised around the presumption of non-financial measures being a causal driver of future financial performance. Nevertheless, the claimed existence of cause-and-effect between non-financial measures and financial performance displays more resemblance to a brute fact than a stylized fact; it is therefore of interest to ask how valid is this assumption? This paper systematically reviews the positivistic management accounting literature for concrete and consistent causal evidence on the association between non-financial measures and financial performance and our findings give rise to concern. The analysis evidenced an unorganised, unstructured and inconsistent body of empirical evidence on the presumed causal relationship between non-financial measures and financial performance. The assumption of non-financial measures being a causal driver financial performance therefore remains unconfirmed and questionable.

Keywords: Causality, Cause-and-effect; Contemporary performance measurement; Non-financial measurement, financial performance

1. Introduction

The Balanced Scorecard (BSC) was one of the first performance measurement conceptualisations that complemented the use of financial measures with non-financial measures (Kaplan & Norton, 1992) on the argument that non-financial measures are leading indicators of future financial performance (Kaplan & Norton, 1996). This argument rested on the assumption of a cause-and-effect relationship to exist between these two types of performance measures and it transformed the BSC from a traditional feed-backward system to a feed-forward system with the ability to predict and control future financial performance (De Haas & Kleingeld, 1999; Malina, Nørreklit, & Selto, 2007).

The introduction of the BSC led to a fundamental change in the perception of performance measurement from traditionally being considered reactive to now proactive (Lueg & Nørreklit, 2013; Micheli & Mari, 2014; Otley, 1999). It also resulted in the BSC becoming the dominant performance measurement system (PMS) in both practice and theory, and, today, contemporary performance measurement always comprises the use of non-financial measures resting on the presumption of these being a causal driver of future financial performance (Franco-Santos, Lucianetti, & Bourne, 2012; Hoque, 2014).

The importance of causality in performance measurement theory has been described as *"Causal relationships among performance management models are an important design criterion and feature of a successful customer loyalty strategy. Conversely, a performance management model without valid causal relations is ineffective or counterproductive to communication and motivation"* (Crosby & Sheery, 2006, p. 13). It is therefore not surprising when the introduction of causality in performance measurement theory has by some been termed as one of the most important and fundamental changes in the history of measuring business performance (Hoque, 2014; Kasperskaya & Tayles, 2013). In the end, it resulted in the presumption that non-financial measures were expected to contain a higher informational value than financial measures when considering long-term performance of an organisation (Atkinson, 2006; Banker, Potter, & Srinivasan, 2000; Barnabè & Busco, 2012; Luft, 2009; Said, Hassabelnaby, & Wier, 2003).

However, the academic discourse on causality in performance measurement literature exhibited no features of a stringent or consistent definition (Janeš, 2014; Kasperskaya & Tayles, 2013; Malmi, 2001; Nørreklit, 2000; Nørreklit, Nørreklit, Mitchell, & Bjørnenak, 2012). Instead, causality was a blurred concept and the claimed existence of cause-and-effect relations displayed more resemblance to a brute fact, which is a fact without explanation or concrete evidence of its existence (Anscombe, 1958; Fahrbach, 2005), than a stylized fact (Kaldor, 1961). A stylized fact is, on the other hand, a fact that refers to empirical findings that are so consistent that it can be accepted as true. A factual transformation from a brute fact to a stylized fact, would in practice require that a pattern for the cause-and-effect relationship is found repeatedly in a field, so that the association is likely to be real, even if its exact extent can be debated. It is a transformation

that positivistic researchers are continuously striving towards by searching for generalisable rules for action, i.e. cause-and-effect relations, and it will always be *the* focus of positivistic management accounting research (PMAR) (Chua, 1986; Ittner, 2014; Lachmann, Trapp, & Trapp, 2017; Luft & Shields, 2014; Micheli & Mari, 2014).

This paper aims at systematically reviewing the consistency of the empirical evidence on causality in contemporary performance measurement literature in an attempt to uncover the validity of the causality presumption. The paper analyses the empirical foundation of the argument of causality and, in doing so, provides an answer to the following research question: *How far has PMAR come in providing consistent empirical answers to causal questions(s) of non-financial measures being leading indicators of future financial performance?* By analysing the progress of knowledge, we provide clarity on the validity of the presumed causal relationship between non-financial measures and financial performance, and, to do so, we analyse the normal science and the related exemplars (Kuhn, 1970). The empirical material in the analysis is a carefully selected range of positivistic performance measurement studies from 1996 to 2014. From the selection of articles, we will then judge which articles are exemplars, and these studies are in a particularly focus, as they should represent the pillars of the argument of causality. In essence, the exemplars are the most well-known solutions to the puzzles of normal science (Kuhn, 1970) and should therefore be the strongest evidence on the existence of cause-and-effect relations. If causality is found to be consistent in the analysed papers and particularly in the exemplars, we provide concrete evidence on the existence of cause-and-effect relations between non-financial measures and financial measures, so that the brute fact of causality in contemporary performance measurement can be translated as a stylized fact instead.

This paper differs in approach and contribution from the theoretical discussion on causality in a special issue published Accounting, Organization and Society in 2014. The special issue mainly discussed methodological issues and provided suggestions to improve statistical inferences on claiming causality (Balakrishnan & Penno, 2014; Gassen, 2014; Ittner, 2014; Luft & Shields, 2014; Lukka, 2014; Van der Stede, 2014). Instead, we provide an empirical and theoretical reflection on the possibility of creating meta-laws i.e. cause-and-effect relations or as Vaivio (2008) frames it *eternal constructs*. We contribute to practice by researching the profoundness of the presumption of causality as practice is reliant on our ability to provide theoretical valid PMS.

In conclusion, we find that the empirical evidence for claiming the existence of causality between non-financial measures and financial performance is at best weak, as there is a clear lack of consistency in the empirical results. We are therefore unable to provide the empirical evidence needed for justifying transforming the fact of causality from a brute fact into a stylized fact and, as such, any recommendation of including non-financial measures in PMS resting on the presumption of causality stands on a flimsy foundation.

The structure of the paper is as follows. In section two, the theoretical framework is explained and causality is defined. Section three clarifies the method for selecting the articles. Section four provides a descriptive overview of the selected articles, while section five provides the analysis of the disciplinary matrix. Section six presents a discussion of the findings from the perspective of quantitative and qualitative research. Finally, the last section provides a conclusion to the paper.

2. Theoretical framework

According to Kuhn (1970) the development of science is not an uniform process but has two alternating stages 'normal' and 'revolutionary' science. Normal science resembles a standard cumulative picture of scientific progress, where science is practiced within a single paradigm and models are constructed as solutions to puzzles under the guidance of a theory. Normal science is therefore a stage of 'puzzle solving' where it is the ambition to continuously strive towards bringing theory and facts into closer agreement (Kuhn, 1970). In normal science, if models fail in solving a puzzle, the theory itself is not criticized or blamed for the failure, as anomalies are not by themselves sufficient for a change in paradigm to occur. In other words, researchers do not renounce a paradigm if anomalies appear because they are not to be treated as a counter instance or as a disproving of theory.

To analyse the progress of positivistic research on the identification of cause-and-effect relations in performance measurement research, we need to consider the disciplinary matrix, as it is designed for analysing paradigms (Kuhn, 1970) and we use the disciplinary matrix as an overlay for evaluating the scientific progress.

2.1 The disciplinary matrix: A theoretical framework for analysing knowledge accumulation

A paradigm is a set of ontological and scientific assumptions that construct a framework from which knowledge can be obtained, acted upon, evaluated and developed (Nørreklit, Nørreklit, & Mitchell, 2010). A paradigm includes the most fundamental hypotheses within a scientific field (Kuhn, 1970). The validity of a paradigm is dependent on the inter-relationship of ontology and epistemology, where a specific ontological assumption implies a particular epistemology, which are to safeguard the validity of the knowledge accumulation (Nørreklit et al., 2010). We argue that the presumption of causality between non-financial measures and financial performance constitutes a significant part of the paradigm of contemporary performance measurement research.

According to Kuhn, a paradigm can be decomposed into the disciplinary matrix dimensions (Kuhn, 1970) allowing for an analysis of stability and progress of a paradigm. The disciplinary matrix consists of four dimensions: *symbolic generalisations, metaphysical presumptions, values and exemplars*. Exemplars found the collars of the paradigm and provide

the basis for normal science to be conducted. They are the most convincing and unique results; whose validity is unquestionable and represent the ideal norms for scientific research within the corresponding paradigm. Exemplars provide empirical content to theory (Kuhn, 1970). Symbolic generalisations are about universal laws or principles that are illustrated by exemplars, for example, that "actions equal reaction" (Kuhn, 1970, p. 182). For contemporary performance measurement it could the generalisation that customer satisfaction is a driver of financial performance. Metaphysical presumptions are described as the *belief* in particular models, as for example the belief in causal PMS such as the BSC. Or more generally, the belief that causality is something that exists within the business world, which can be empirically verified and transformed into universal generalisations to be considered when implementing PMSs. Lastly, values are defined as normative conceptions of what forms the qualities of a scientific theory that are applied in the selection of competing theories or paradigms. In other words, it is about predictions. They should be accurate, whatever the margin of permissible error, it should be consistently satisfied in a given field and they should be used to judge whole theories. For contemporary performance measurement values concerns the predication, accuracy and consistency in finding cause-and-effect relations or the relationship between the use of certain PMSs and the financial performance of an organisation.

In the state of normal science, the disciplinary matrix has consensus and is kept fixed allowing for knowledge accumulation through the cumulative generation of puzzle solutions. While, it is unstable and lacks consensus in periods of revolutionary science where anomalies are present (Bird, 2013). The knowledge accumulation process of a paradigm consists of normal science where theory is not being questioned; it is a cumulative progress with no meaning variance, and change is incremental and gradual (Bird, 2013). The paradigm of contemporary performance measurement, within the positivistic research tradition, is concerned with empirically verifying cause-and-effect relations between non-financial measures and financial performance and between contemporary PMS and financial performance, this is what currently can be considered normal science and 'puzzle-solving'. However, to analyse the progress on providing the paradigm with empirical content, we first need to develop and define the concept of causality and what it requires from a scientific method in terms of discovery.

2.2 The metaphysical presumption of causality[5]

To discover causality between two phenomena, one would have to rely on experience and observations as *"it is impossible for us, by any sagacity or penetration, to discover, or even*

[5] This paper draws upon David Hume's definition and understanding of causality developed in his book *"An Enquiry concerning Human Understanding"* from 1748, as it is his criteria, which are usually assumed within theory of science (Edwards, 1972; Slife & Williams, 1995) and also within theory of statistics in social sciences (Angrist & Pischke, 2014; Simon, 1970).

conjecture, without experience, what event will result from it, or to carry our foresight beyond that object which is immediately present to the memory or senses" (Hume, 1975, p. 74). This means that a causal relationship between X and Y cannot be logically reasoned, it can only be empirically observed and substantiated through experiences. Hume continues and argues that to claim a universal law, a causation, one would have to rely on an endless reproduction of the same relationship: *Even after one instance or experiment, where we have observed a particular event to follow the upon another, we are not entitled to form a general rule, or foretell what will happen in like cases; it being justly esteemed an unpardonable temerity to judge of the whole course of nature from one single experiment, however accurate or certain"* (Hume, 1975, p. 74). To claim the existence of a universal law, one would therefore have to rely on an endless reproduction of the same connexion of two phenomena regardless of *time, place and context*, as *"when one particular species of event has always, in all instances, been conjoined with another, we make no longer any scruple of foretelling one upon the appearance of the other, and of employing that reasoning, which can alone assure us of any matter of fact or existence. We then call the on object, Cause; the other, Effect. We suppose that there is some connexion between them; some power in the other, by which it infallibly produces the other, and operates with the greatest certainty and strongest necessity"* (Hume, 1975, pp. 74-75).

After describing what causality implies, he provides his twofold definition that consists of an external and internal impression. First, *"A cause is an object followed by another, where all objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second had never existed"* and, second, *"A cause is an object followed by another, and whose appearance always conveys the thought to the other"* (Hume, 1975, pp. 76-77). The first part provides the relevant external impression, while the second captures the internal impression, which is our awareness of being determined by custom to move from cause to effect, and only together do they capture all the relevant impressions of a causal relationship. Therefore, to have knowledge of a cause-and-effect relationship, would be to have perfect knowledge on their connexion, and this would allow one to predict and control future events by their causes *"For surely, if there be any relation among objects which it imports to us to know perfectly, it is that of cause and effect. On this are founded all our reasonings concerning matter of fact and existence... The only immediate utility of all sciences, is to teach us, how to control and regulate future events by their causes."* (Hume, 1975, p. 76).

Hume's definition of causality implies 'universality' in generalisation, in other words, a cause-and-effect relationship would be an eternal construct (Angrist & Pischke, 2014; Hubbard & Lindsay, 2013; Vaivio, 2008). Cause-and-effect relations are built upon notions of regularity, explanation, predictability and control and knowledge accumulation occurs when cause-and-effect relations are founded upon earlier cause-and-effect relations i.e. putting layers upon layers of knowledge (Wright, 1994).

PMAR focuses on causal explanations. For example, do certain activities *drive* overhead costs? Does the adoption of a balanced scorecard *improve* performance? It is therefore about drawing inferences from a sample of specific observations to the general (Lachmann et al., 2017). In this sense, PMAR implicitly adopts the Humean notion of causality, which implies an ambition to discover universal laws (cause-and-effect relations) through scientific inference (i.e. hypothetico-deductive method) (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2014; Micheli & Mari, 2014). For contemporary performance measurement the identification of cause-and-effect relations result in an ability to predict and control for future events (financial performance) by their causes (non-financial measures) (De Haas & Kleingeld, 1999; Lueg & Nørreklit, 2013). A discovery of cause-and-effect relations is therefore also of impeccable value to practice and a search for such relationships is therefore of course inevitable for PMAR (Ittner, 2014; Lachmann et al., 2017; Micheli & Mari, 2014).

2.3 The scientific method of casual discovery

The scientific method for finding causality is controlled experimental designs, which unfortunately is a tool unavailable to researchers within social sciences (Evans, Feng, Hoffman, Moser, & Stede, 2015; Goodman, Fanelli, & Ioannidis, 2016; Wold, 1954). In controlled experiments, researchers can allow for one or more variables to be under their control, and for suitable chosen values the researchers are able to observe the values of one or more other variables whose variation is of interest. When an experiment reveals that an observed variable varies systematically, as the controlled variables are allowed to vary, just then can a researcher claim causality. Another important feature of controlled experiments is that it allows for other researchers to replicate the experiment and hence reproduce the causal relationship; it is herein the supremacy of controlled experiments in uncovering causality lies (Wold, 1954). As, to identify a causal relationship, it ultimately requires an endless reproduction. However, studies that replicate previous findings in social sciences are few and far between (Dyckman & Zeff, 2014; Goodman et al., 2016), but reproducibility is a cornerstone of the hypothetico-deductive method, as if an empirical finding is to be considered a 'fact', other researchers must be able to observe it, thus strengthening the credibility of the 'fact' (Kane, 1984).

In PMAR the method of the hypothetico-deductive method along with null hypothesis testing (NHST) is the most consistent model of scientific explanation (Chua, 1986) and it is often referred to as *the* scientific method (Lachmann et al., 2017; Lindsay, 1994). According to Chua (1986), the hypothetico-deductive account of scientific explanation has two main sequences. First, it is about searching for universal laws or principles and thereby to explain an event is to present it as an instance of a universal law. Second, it is about being able to predict and control for future events by their causes, as when an event is explained its occurrence can be deduced from the premises. It follows that knowing the premises before the event happened would enable a prediction that it would happen. It would also enable steps to be taken to control the occurrences

of the event. The search for universal regularities and causal relationships is ubiquitous for PMAR.

These assumptions for the scientific 'explanation' influence the choice of research methods, as research reports begin with a statement of hypotheses followed by a discussion of empirical data and concluded with an assessment of the extent to which the data "supported" or "confirmed" the hypotheses (Chua, 1986). Data collection and analysis is therefore focused on the "discovery" of rigorous and generalisable relationships i.e. cause-and-effects. This means that PMAR neglect "soft" methods such as the case study.

Unfortunately, the ability of social sciences to transform empirical findings into cause-and-effect relations is being questioned as there is an ongoing and widespread replication crisis in natural and social sciences (Baker, 2016; Camerer et al., 2016; Gelman, 2015; Goodman et al., 2016; Maniadis, Tufano, & List, 2014; McNutt, 2014). It is therefore important that causal studies justify for the following: *"(i) a simple structure that you can see through to the data and the phenomenon under study, (ii) no obvious plausible source of major bias, (iii) serious efforts to detect plausible biases, efforts that have come to naught, and (iv) insensitivity to small and moderate biases"* (Gelman, 2011, p. 965). Studies claiming the existence of cause-and-effect relations must ensure that the relations are theoretically plausible, as mindless statistical inference despite it being significant are unlikely to result in the discovery of causal relations (Gigerenzer, 2004; Gigerenzer & Marewski, 2015). David Hume also cautioned researchers to attain a certain amount of mitigat3ed scepticism, e.g. to be *tentative, modest, reserved, cautious and at all-time being sceptical* (Hume, 1975). It is essential that an argument of causality does not rely on a bright-line rule i.e. $p < 0.05$, as a statistical conclusion does not immediately become 'true' on one side of the divide and 'false' on the other side (Chawla, 2017; Fisher, 1956; Nuzzo, 2014; Wasserstein & Lazar, 2016).

In the light of this discussion, we have provided a theoretical argument for the relevance of studying the consistency of causal claims; if we are to transform the notion of causality in contemporary performance measurement from a brute fact to a stylized fact. Inconsistent evidence on causality would, on the other hand, raise doubt to whether the found relationship truly was causal and not just a statistical significant correlation.

3. Method: Selection of articles and a qualitative investigation of causality

For the purpose of this paper, 11 prestigious accounting journals are selected and they provide a large and representative sample of the empirical research conducted on identifying causal relations in contemporary performance measurement (see appendix A).

The search method for finding relevant articles was inspired by Papaioannou, Sutton, Carroll, Booth, and Wong (2010) and follows the approach of conducting a systematic literature review. We employ conventional subject searching, reference list checking and citation search on

Web of Science Two different databases are used to increase the chances of all relevant articles being found, and to ensure that the databases support the selected time-period for the study, namely papers on identifying causal relations within performance measurement research from 1996-2016.

The following search code was conducted for each of the selected journals: "ISSN(...) and AB(nonfinancial measures or non-financial measures)" in the databases ABI Inform and Business Source Complete. A broad keyword terminology is employed, as it is often difficult to find precise keywords to search for (Papaioannou et al., 2010).

The search code provided 44 articles from ABI/Inform and 34 articles from Business Source Complete; hence, in total 78 articles were identified. The title, abstract and subject for each article were browsed in order only to include relevant articles; this left us with 30 articles. For these 30 articles, the introduction, methodology and conclusion were subsequently skimmed in order to determine their relevance and to exclude those not relevant for this study; this left us with 15 articles (see appendix B). These 15 articles were subject to reference list checking and citation search in order to heighten the internal validity of the analysis, but no further articles were identified.

The remaining articles were categorized by author, key issues addressed, research method, data, key findings, limitations, and if causality was defined. A strategy often employed by other literature reviews within management accounting research (Chenhall & Smith, 2011; Hoque, 2014; Scapens & Bromwich, 2001; Shields, 1997b). The categories were chosen due to them representing the core elements of an article. The analysis of the papers will be based on the disciplinary matrix framework and conducted as a qualitative investigation of the knowledge accumulation by PMAR on the uncovering of consistent empirical evidence on the existence of causality in contemporary performance measurement.

4. A descriptive overview of the articles in analysis

This section provides a preliminary impression on the consistency of empirical evidence on cause-and-effect relations in the 15 articles.

Table 1, lists the authors, journal, year published, methods, findings, citations and exemplars and the analysis includes nine longitudinal studies, four cross-sectional studies and two quasi-experimental studies. We use citations to identify exemplars, as these studies should be the most convincing; whose validity is unquestionable and represent the ideal norms for scientific research within the corresponding paradigm. It would therefore be reasonable to assume that such studies are the most well-cited. These studies are the pillars of normal science and should provide an empirical argument to the presumption of causality. The exemplars are represented by two longitudinal studies, two cross-sectional studies and two quasi-experimental studies.

Table 1. A descriptive list of the articles on authors, year, journal, type of study, findings and citations

Author (year)/journal	Type of study	Causal relations	Citations*	Exemplar
Perera, Harrison and Poole (1997)/AOS	*Cross-sectional*	*No evidence*	*462*	Yes
Ittner, Larcker and Randall (2003)/AOS	*Cross-sectional*	*No evidence*	*1017*	Yes
Wiersma (2008)/AOS	Longitudinal	Mixed evidence	44	No
Hoque (2005)/BAR	Cross-sectional	Contextual	128	No
Davis and Albright (2004)/MAR	*Quasi-experiment*	*Significant*	*646*	Yes
Ittner and Larcker (1998)/JAR	*Longitudinal*	*Mixed evidence*	*1710*	Yes
Banker, Potter and Srinivasan (2000)/TAR	*Quasi-experiment*	*Significant*	*1045*	Yes
Nagar and Rajan (2001)/TAR	Longitudinal	Significant	161	No
Sedatole (2003)/TAR	Longitudinal	Contextual	59	No
Dikolli, Kinney JR and Sedatole (2007)/CAR	Longitudinal	Significant	36	No
Banker and Mashruwala (2007)/CAR	Cross-sectional	Contextual	81	No
Malina, Nørreklit and Selto (2007)/CAR	Longitudinal	No evidence	102	No
Said, HassabElnaby and Wier (2003)/JMAR	*Longitudinal*	*Mixed evidence*	*409*	Yes
Smith and Wright (2004)/JMAR	Longitudinal	Significant	219	No
Dikolli and Sedatole (2007)/JMAR	Longitudinal	Contextual	40	No

*Found on Google Scholar - 06/09-2016

All papers used a hypothetico-deductive approach along with NHST as a method. The papers approached the topic of causality in contemporary performance measurement differently. Some papers studied specific cause-and-effect relations between non-financial measures and financial performance, while other studies approached causality on a system level, i.e. if the use of non-financial measures on a general level lead to higher financial performance. Lastly, some studied how contextual factors mediated the correlation between non-financial measures and financial performance. These studies therefore tried to uncover if cause-and-effect relations where subject to the influence of contextual variables or mediators. However, there was no replications present amongst the 15 studies, as all studied different hypothesised relations. The ambition of these studies is to discover empirical relationships between non-financial measures and financial performance that are rigorous and generalisable so that future events can be controlled and predicted by their causes (Chua, 1986). In other words, it is to search for universal laws or principles, i.e. causality, that can be generalised to the business world (Ittner, 2014; Lachmann et al., 2017; Micheli & Mari, 2014).

It is surprising that only one of these studies discussed, defined and concluded upon whether the significant correlations where actual causations (Malina et al., 2007), while, two studies explicitly contradict their own ontological and epistemological assumption on causality, as they argued that causality could be bi-directional, reverse or reciprocal (Perera, Harrison, & Poole, 1997; Wiersma, 2008). A bi-directional, reverse or reciprocal causal relationship would be in conflict with the Humean definition and criteria for causality. The potential for such contradictions to occur illustrates that causality, in the field of management accounting, is a concept which largely remain undiscussed and instead it is implicitly assumed or operationalised when findings are to be generalised. This is also why statistical research argues that when

researchers claim causality, it reflects their mental representation, as it is the language used, which determine the reliability of the judgement on which the analysis of the significant correlation so crucial depends. If it is to be understood as a cause-and-effect relation (Pearl, 2010). It is important that studies on correlations with the obvious ambition of claiming causality, must actively reflect on the plausibility of correlations being cause-and-effect relations, which is something that the papers in the analysis appears to be lacking.

5. Analysis

In the following, we provide our qualitative investigation of the analysed papers in assessing the state of knowledge accumulation and the strength of the metaphysical presumption of causality. The analysis is structured around the four dimensions of the disciplinary matrix by first analysing the exemplars, then the symbolic generalisations and metaphysical presumptions, and lastly the value dimension.

The exemplars are the articles that found the collars of a paradigm and provide the basis for normal science to be conducted. These studies should be most convincing and unique, while pertaining an unquestionable validity; i.e. represent an ideal norm for scientific research. In addition, exemplars are intended to create tacit knowledge of a theory (Kuhn, 1970), which in this case is the evidence on specific cause-and-effect relations while the overarching theory is the tacit assumption of causality to exist within contemporary performance measurement. By analysing the exemplars, we provide evidence on the development of tacit knowledge and whether the overarching theory is supported, which is a necessity for normal science to be sustainable (Kuhn, 1970). It is essential that these articles reach similar conclusions in terms of evidencing causality, as else it would be impossible to form universal laws or principles from the exemplars i.e. symbolic generalisations.

In table 1, we have highlighted (in italic) the articles judged to be the exemplars on causality between non-financial measures and financial performance in contemporary performance measurement theory. These papers are in chronologically order Perera et al. (1997), Ittner and Larcker (1998), Banker et al. (2000), Ittner, Larcker, and Randall (2003), Said et al. (2003) and (Davis & Albright, 2004), and they represent more than 86% of all citations (counted in Google Scholar). Table 2 lists all of the hypotheses related to identifying cause-and-effect relations between non-financial measures and financial performance, and table illustrates that the exemplars provide inconsistent empirical evidence on causality, as some claimed to have found cause-and-effect relations, while others claim the opposite, and still others provide weak to modest empirical evidence for the claim of causality. It raises a preliminary concern for the ability of the exemplars to provide a foundation for symbolic generalisations i.e. the creation of universal laws.

Table 2. An overview of the findings on each hypothesis from all of the exemplars.

	Hyptheses	Significance
Perera, Harrison and Poole (1997)	H2: Increasing use of non-financial performance measures is associated with enhanced performane for firms pursuing customer-focus in manufacturing strategy, as proxied by the implementation of AMP and AMT.	No evidence
Ittner and Larcker (1998)	H1: Are current satisfaction levels for individual customers associated with changes in their future purchase behavior and firm revenues?	Modest support
	H2: To which extent do business-unit customer satisfaction mesures predict future accounting performance and number of customers?	Modest support
	H3: Does the stock market view customer satisfaction as a forward-looking performance indicator?	Some evidence
Banker, Potter and Srinivasan	H: Does increased emphasis on customer satisfaction lead to an increase in revenue and operating profit?	Yes
Ittner, Larcker and Randall (2003)	H1: Organizational performance is positively associated with the extent to which the firm measures and uses information related to a diverse set of financial and non-financial performance measures.	Weak support
	H4: Organizational performance is positively associated with the use of balanced scorecards, economic value measures, and causal business models?	No evidence
Said, HassabElnaby and Wier (2003)	H1: Firms that use a combination of non-financial and financial measures perform better contemporaneously than firms that use financial measures alone.	Mixed
	H2: Firms that use a combination of non-financial measures and financial measures perform better prospectively than firms that use financial mesures alone.	Yes
Davis and Albright	H: Is changes in financial performance acheived with the implementation of a BSC program?	Yes

Hypotheses not analysing a relationship between non-financial measures and organisatonal performance have been removed from the analysis due to irrelevance.

5.1 Exemplars: What do they tell us on causal questions?

The analysis of the exemplars is structured after the different methods applied by the papers, so that cross-sectional studies are first analysed, then longitudinal studies and lastly, quasi-experimental studies.

The study by Perera et al. (1997) analyse cross-sectional data and tested whether an increase in non-financial measures would be associated with enhanced performance for firms pursuing a customer-focused strategy based on the components of AMP[6] and AMT[7]. The study tested various interaction regression models and only in the case of testing the interaction between non-financial measures and AMP did the model yield a positive significant result with a corresponding R square of 0.06. In the end, Perera et al. (1997, p. 569) concludes that *"The study was not able to find a consequential link to organizational performance... However, in the absence of a finding for costumer-focus generally, and with no convincing argument for the*

[6] AMP: Advanced Management Practices
[7] AMT: Advanced Manufacturing Technology

technology component on its own, the results must be seen as failing to support the link to performance".

A cross-sectional study by Ittner et al. (2003) first investigates whether organisational performance is positively associated with the extent to which a firm uses a diverse set of financial and non-financial performance measures and second, the study analyses the relation between organisational performance and the use of various PMSs. They find that non-financial measurement is insignificantly correlated with ROA, sales growth and three-year stock return; however, it is significantly correlated with one-year stock returns with an adjusted R square of 0.088 and a coefficient size of 0.0843. This study casts doubt on whether a greater measurement diversity of non-financial measurements is associated with higher organisational performance. Whether the use of PMSs is associated with an increase in performance, Ittner et al. (2003) finds the use of Economic Value Added to be insignificant associated with organisational performance measures such as ROA, sales growth, one-year stock return and three-year stock return. The Balanced Scorecard was also found to be negatively significant associated with ROA [-0.0117, t = -2.787, $p < 0.01$] and insignificant in relation to sales growth, one-year stock returns and three-year stock returns. On the other hand, business modelling was found to be significantly and positively associated with ROA [0.0078, t = 1.67, $p < 0.10$], while also insignificant in relation to sales growth, one-year stock return, and three-year stock return. The study therefore concludes, *"these results provide little support for the hypothesis that the use of these measurement alignment techniques influence economic performance"* (Ittner et al., 2003, p. 736).

The longitudinal study by Ittner and Larcker (1998) examined the causal question whether customer satisfaction measures provide economic value to an organisation. The study approached the question from three perspectives: First, are customer satisfaction measures leading indicators of accounting performance? Second, is the economic value of customer satisfaction reflected in contemporaneous accounting book values? Third, does the release of customer satisfaction measures provide new or incremental information to the stock market? (Ittner & Larcker, 1998, p. 1). The study found that the Customer Satisfaction Index (CSI) was significantly positively correlated with customer retention [0.002, t = 6.16, $p < 0.01$, adj. R2 = 0.021], revenue [19.464, t = 4.92, $p < 0.01$, adj. R2 = 0.049] and revenue change [0.003, t = 5.74, $p < 0.01$, adj. R2 = 0.013]. Despite being significantly and positively correlated, the low explanatory power along with low coefficients suggests that the relationship is rather weak and that the that are other factors which are more important. Next, the study found that CSI on business-unit level was significantly associated with revenue [2.089, t = 2.35, $p < 0.05$] and the number of business and professional customers (B&P) [0.489, t = 1.78, $p < 0.10$], while CSI was not statistically significant with expenses, margins, ROS, or retail customers. These results indicate that branches with higher satisfaction scores had higher revenue per customer and an indirect effect on accounting performance through attracting new customers. However, the findings also

indicated that many of the accounting gains mainly appeared to come indirectly through new growth in customers instead of increased profits from existing customers. Lastly, the study found that the customer satisfaction constructs (ASCI) for 1994 and 1995 were significantly and positively associated with assets [1.73, $p < 0.01$ and 2.19, $p < 0.01$] and liabilities [-1.77, $p < 0.01$ and -2.25, $p < 0.01$], implying that customer satisfaction measures provide insight into firm value which is not reflected in current accounting book values. From these results, the study concludes that they found that *"modest support for customer satisfaction measures are leading indicators of customer purchase behaviour (retention, revenue, and revenue growth), growth in number of customers, and accounting performance (business-unit revenues, profit margins, and return on sales). We also found some evidence that firm-level customer satisfaction measures can be economically relevant to the stock market but are not completely reflected in contemporaneous accounting book values"* (Ittner & Larcker, 1998, p. 32). The study also indicated that customer behaviour and financial results were only relatively constant over broad ranges of customer satisfaction levels, but changed after satisfaction moved through various threshold values and diminish at higher satisfaction levels. This indicates that the relationship is not causal, as a cause-and-effect relationship according to its definition is stable in magnitude.

The study by Said et al. (2003) examines through longitudinal data whether firms that use a combination of non-financial and financial measures perform better, contemporary or prospectively, than firms using financial measures alone. The study found that non-financial measures were insignificantly correlated with contemporaneous financial performance while they were positively correlated with prospectively financial performance [0.053, $t = 3.19$, $p < 0.01$]. In addition, the study found that non-financial measures were significantly correlated with market-adjusted stock returns. These results suggest that non-financial measures appear to be leading indicators of future accounting-based performance, however, they yield no impact on current accounting-performance as measured by return on assets. The study concludes, *"Although we find some evidence for future accounting-based performance, the overall evidence on financial and non-financial measures' impact on accounting-based performance is mixed"* (Said et al., 2003, p. 217).

The study by Banker et al. (2000) is a quasi-experimental study which tests whether an non-financial measures are leading indicators of financial performance. The study examined the two customer satisfaction measures of likelihood of return (LRETURN) and customer complaints (COMPLNTS) in relation to the financial measures of revenue, costs and profits. The study found a positive significant correlation between LRETURN and revenue [35.71, $t = 4.37$, $p < 0.01$] and profit [19.51, $t = 3.307$, $p < 0.01$], while it exhibited no significant correlation with costs. COMPLNTS exhibited no significant correlation with any of the three financial measures. On the other hand, the study found no significant correlation between LRETURN or COMPLANTS and current revenues or profits, which according to the study indicates that

customer satisfaction impacts future rather than current performance. The study concludes, *"We documented that customer-satisfaction measures... are significantly associated with future financial performance as measured by revenues and operating profits, but not with operating costs... The positive association between future revenues and current nonfinancial performance is mainly driven by occupancy (volume effect) as opposed to room rates (price effect)"* (Banker et al., 2000, p. 89).

The last exemplar is a one-year quasi-experimental study by Davis and Albright (2004) which investigates whether bank branches implementing the BSC outperform bank branches within the same banking organization on financial performance. The study used a Wilcoxon rank test to examine if there was a difference in performance between the experimental and control division branches. The study found a significant difference in the expected direction between average performance of the experimental and control division branches [$z = 1.826$, $p < 0.034$]. The study further illustrated that the increase in performance occurred subsequent to implementation of the BSC in the experimental branches [$z = 2.205$, $p < 0.014$], while control branch performance did not improve. Davis and Albright (2004, p. 150) concludes from these findings that *"We provide evidence supporting the proposition that the BSC can be used to improve financial performance; the findings indicate branches in the BSC group outperformed non-BSC branches on a common composite financial measure. The research method and design of this study allow for a causal statement concerning the association between BSC implementation and financial performance improvement"*.

To summarize on the findings, we conclude that: On the question whether customer satisfaction is a causal driver of financial performance, Ittner and Larcker (1998) found the correlation to be non-linear by being diminishing at high satisfaction levels and only affecting financial performance when moving through certain thresholds. On the other hand, Banker et al. (2000) found that the customer satisfaction measure of 'likelihood of return' was correlated with revenue while not related to costs, however, only in relation to future financial performance as the measured exhibited no relation to current financial performance and the driving factor was volume effects and not price effects. In addition, the study found that complaints were not related to financial performance neither in terms of revenue nor in terms of costs.

On the question whether the use of non-financial measures was a causal driver of financial performance, Perera et al. (1997) concluded that they were unable to find any consequently link between a customer-focused strategy and financial performance. Corroborating these results, Ittner et al. (2003) found that a non-financial measurement focus was insignificantly related with a diverse set of financial measures e.g. ROA, sales growth and three-year stock return, while it only exhibited a very weak correlation with one-year stock return as the seize of the coefficient was 0.0843 and the R square for the model was 0.088. On the contrary, Said et al.

(2003) found non-financial measures to be related to prospectively financial performance while unrelated to contemporaneous financial performance.

Lastly, on the question whether contemporary performance measurement models were a causal driver of financial performance, Ittner et al. (2003) found that the use of economic value was insignificant, while the BSC was negatively related with ROA and insignificant with other financial performance measures. Finally, the study found that business modelling was weakly related with ROA with a coefficient of 0.0078, while insignificant with other financial measures. In contrast, Davis and Albright (2004) found that the use of the BSC in bank branches was related to a significant difference on average performance, however, the study did not provide any information on the size of the effect.

Overall, the exemplars tried to identify three cause-and-effect relations: Is customer satisfaction a driver of financial performance? Are non-financial measures leading drivers of financial performance? Are contemporary PMSs drivers of financial performance? Considering that only *"when one particular species of event has always, in all instances, been conjoined with another... We then call the one object, Cause; the other, Effect."* (Hume, 1975, p. 74), it is not possible to conclude on any of these supposed cause-and-effect relations. The evidence is unclear on all three relationships, as the evidence is contradictory, inconsistent, imprecise and produces weak relations when considering the effect sizes. This is not to say that there is no relationship between non-financial measures and current and future financial performance; the evidence is just not able to make it causal from a Humean perspective. It is more likely to be a relationship that is *dependent* on time, place and context and hence not universally generalisable. Other types of relationships do exist; Hume describes them as resemblance and contiguity (relationship in time or place) (Hume, 1975), however none of those bring forth an ability to predict, control, and regulate future events by their causes.

According to Kuhn (1970), inconsistency is problematic for knowledge accumulation to occur within a paradigm and, as such, the creation of tacit knowledge that could be generalised to practice is non-existent or at least unreliable. From a Kuhnean view, we are therefore unable to build layers upon layers of knowledge if the anomalies of the underlying relations remain unsolved i.e. notions of regularity, explanation, predictability and control cannot be justified.

5.2 Symbolic generalisations: The meta-laws of contemporary performance measurement

Contemporary performance measurement is developed on the notion that non-financial measures are leading indicators of financial performance resting on the presumption of discoverable cause-and-effect relations which would transform performance measurement from being reactive to proactive thereby providing the ability to control and regulate future events [financial performance] by their causes [non-financial performance].

The academic consensus on the presumption of causality between non-financial measures and financial performance is in the disciplinary matrix described as symbolic generalisation (Kuhn, 1970). A symbolic generalisation is therefore a meta-law of universal generalisability like "actions equals reaction" (Kuhn, 1970, p. 182) or as Hume describes it *"we either mean that this vibration is followed by this sound, and that all similar vibrations have been followed by similar sounds: Or that this vibration followed by this sound, and upon the appearance of one the mind anticipates the senses, and forms immediately an idea of the other*" (Hume, 1975, p. 77). For example, in natural science a symbolic generalisation could be Newton's laws of motion.

For contemporary PMSs, such a law could be that non-financial measures drive long-term financial performance or more particular that customer satisfaction drives short and long-term financial performance; it would be a meta-law of contemporary performance measurement. However, the exemplars analysed in the previous section are unable to provide such a symbolic generalisation with any consistent empirical content. Instead, the exemplars evidence that the auxiliary and theoretical assumption of causality appears to be absent in practice or at least not universal generalisable to practice. It indicates that the relationship between these phenomena is of another nature than causal, as these relationships appear to depend on time, place and context.

5.3 Metaphysical presumptions: Is causality to be found?

Kuhn described metaphysical presumptions as the *belief* in particular models. For example, in natural science a metaphysical presumption could be that atoms are like 'billiard balls', or light is like a wave or 'particles'; it provides the preferred analogies.

If we relate this to contemporary performance measurement, then it could be the belief in particular PMSs such as the BSC or more generally the belief that non-financial measures are causally related to financial performance and that it can be generalised to any practical setting in any time horizon. Reflected in PMAR trying to answer questions such as 'do certain activities drive overhead costs'? Or 'does the implementation of a BSC improve performance'? (Ittner, 2014). However, there are is almost no discussion of theory. Papers lacked a compelling theoretical reflection to why an association between non-financial measures and financial performance should be of a causal nature, which is considered to be important when arguing for the existence of something more than just a correlation (Wasserstein & Lazar, 2016). In addition, there is no testing of an overarching theory but instead fragmented hypotheses on how various non-financial measures or contemporary PMS should be related to financial performance. In the end, the analysis of the exemplars indicates that the belief in these analogies or cause-and-effect relationships such as the belief in quality or customer satisfaction as a *causal* driver of financial performance is based on inconsistent and inaccurate evidence, and it begs the question of whether these beliefs/analogies should be generalised to practice.

5.4 Values: Is NHST a universal method for scientific inference?

Values concern predictions; they should be accurate, whatever the margin of permissible error, they should be consistently satisfied in a given field, and they should be used to judge whole theories. They should first and foremost permit puzzle-formulation and solution (Kuhn, 1970, pp. 184-185). We look at the scientific method for studying and identifying causality in social science as the method is about making accurate prediction and solving puzzles.

For all six exemplars, the scientific method for uncovering cause-and-effect relations has been the hypothetico-deductive method, i.e. null hypothesis testing (NHST), which is also the case for the rest of the papers in the systematic review. In general, this is the approach of PMAR (Chua, 1986; Lachmann et al., 2017; Shields, 1997a).

However, recent controversies have cast doubt on the reliability of NHST as a statistical tool for claiming causality. It has been evidenced that statistical inferences are proven to be unreliable, lacking replicability, and facing emerging problems with generalising 'lab' findings to a real-world setting (Camerer et al., 2016; Gelman & Loken, 2014; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). It appears that a set of questionable research practices (QRPs) are distorting the hypothetico-deductive method in favour of researchers' own hypothesis, which upends the probability for a significant result being a false-positive one (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Ioannidis, 2005; Simmons et al., 2011).

Accuracy of prediction is not only about the significance of a hypothesis, as a focus on significance might lead to the production of spurious relationships; it is just as important to consider the magnitude of the significant relationship (Wasserstein & Lazar, 2016). As such, it is concerning when the exemplars judge causality solely on the significance, i.e. $p < 0.05$, instead of a collective consideration of the plausibility of the relationship based on a composite of statistical measures e.g. power analysis, confidence intervals and effect sizes. P values do not imply the likelihood of a correlation to be replicable (Wasserstein & Lazar, 2016). It is unfortunate when exemplars create precedence for a 'bright-line' rule ($p < 0.05$) for which findings becomes scientific important or, in other words, *causal* (Gigerenzer & Marewski, 2015). Magnitude is just as important when judging something to be economic and scientific important (Halsey, Curran-Everett, Vowler, & Drummond, 2015). In consequence, it is unclear whether the embedded values in the exemplars and disciplinary matrix are able to provide accurate predictions, and thereby provide successful puzzle solutions.

To sum up, in Kuhn's most generic usage, a paradigm is what the members of a 'scientific community', and they alone, share. This community will to a remarkable extent have absorbed the same literature and drawn similar lessons from it, and a disciplinary matrix consists of shared exemplars that a discipline or scientific community considers to be 'good science'. Thus, a paradigm is established on a common set of beliefs and methods through their ability to conduct

convincing and consistent puzzle solving (Kuhn, 1970). From this perspective, the findings from the analysis question the presumption of causality and the existence of cause-and-effect relations, as the empirical evidence in our perception are too vague and uncertain to provide the foundation needed for a paradigm to formalise.

The analysis evidences an unorganised and unstructured body of empirical evidence on cause-and-effect relations providing divergent results on the existence of causality in contemporary performance measurement. The causal question of non-financial measurement driving financial performance remains unanswered. This does not mean that non-financial measurement does not partake in the financial performance of a firm, it implies that currently it is not possible to claim this association to be causal.

6. Discussion

When research fails in puzzle solving and while in the stages of normal science, the theory itself will not be criticized or blamed for the failure, as anomalies by themselves are not sufficient for a paradigm to chance (Kuhn, 1970). Researchers will not renounce a paradigm if anomalies appear because they are not treated as counter instances, or in other words, disproving the theory. Though in the vocabulary of science that is what anomalies are (Kuhn, 1970). It is not only quantitative researchers who are preoccupied by the challenges of causality in social sciences, we will therefore divide our discussion into two sections by first providing quantitative perspective and then a qualitative.

6.1 Quantitative research on causality

From the ontology of positivistic research, a theory is to be judged on its ability to predict the phenomenon it is intended to 'explain' (Friedman, 1953; Zimmerman, 2001). The study found a conflicting body of evidence, indicating that non-financial measures in some cases are related to financial performance (Ittner et al., 2003; Perera et al., 1997; Said et al., 2003), sometimes unrelated (Perera et al., 1997), at other times the evidence is highly mixed (Ittner & Larcker, 1998), exhibiting no relation to costs (Banker et al., 2000), and when measured on a system level, there is even a possibility for a deterioration of financial performance (Ittner et al., 2003). From this perspective and the findings in this study, it is reasonable to question, from a positivistic standpoint, if we have any 'theory' on causality in contemporary performance measurement. However, the study did not find sufficient evidence on an association between non-financial measures and financial performance to claim it a cause-and-effect relation. As such, it seems reasonable to conclude that the exemplars are unable to provide an accurate and reliable prediction of the phenomenon they intend to explain or causal questions they aim at answering.

According to Kuhn (1970), a paradigm is dependent on the ability of conducting convincing and consistent puzzle solving. However, the PMAR analysed in this paper on causality in contemporary performance measurement is unable to develop what could be considered a

consistent and reliable theory with 'proved' hypotheses that could support practice in the development and implementation of PMSs, which is argued to be a central purpose of management accounting research (Baldvinsdottir, Mitchell, & Nørreklit, 2010; Micheli & Mari, 2014; van der Meer-Kooistra & Vosselman, 2012). Causality is what should render PMS proactive of the future, however, our findings indicate that this claim is not well-argued or well-evidenced and we therefore question the reliability of this presumption. As such, it is not surprising when often implement contemporary PMSs without explicit [presumed] causal relationships (Albertsen & Lueg, 2014; Gates, 1999; Greiling, 2010; Ittner et al., 2003).

The anomalies of insignificant relationships between non-financial measures and financial performance has evidently led to a reflection on causality and methods in PMAR as a special issue on causality in Accounting, Organization and Society (AOS), reflects on the technical aspects of scientific methods in making causal inferences (Balakrishnan & Penno, 2014; Gassen, 2014; Ittner, 2014; Luft & Shields, 2014; Lukka, 2014; Van der Stede, 2014). The special issue focused on the value dimension of the disciplinary matrix i.e. the scientific method for making causal inferences, with the result of suggestion more advanced statistical models to overcome the issues of making causal claims. However, a continuous advancement towards more advanced and complex statistical models has previously been argued to be a defensive strategy for protecting a paradigm from critique, as potential critics would lack the technical know-how of these models (Bamber, Christensen, & Gaver, 2000).

When we isolate a particular feature or association of the world, a cause-and-effect, is it then an eternal truth? Or a local truth? According to Gelman (2014), the uncertainty and variation of the social world provide two reasons for why neither is likely to be the case. The issue of uncertainty relates to the issue of studying small effects because it is very possible that a large proportion of statistical significant findings are in the wrong direction as well overestimating the magnitude of the underlying effect (Gelman & Tuerlinckx, 2000). If we consider variation, then if a finding is in fact 'real', in the sense of having the same sign as the corresponding comparison in the population, it might be different in other populations or scenarios. In short, *"an estimated large effect size is typically too good to be true, whereas a small effect could disappear in the noise"* (Gelman, 2014, p. 641). A solution to these issues is replications, which unfortunately is a rare type of research in social sciences but common to natural sciences (Goodman et al., 2016). However, it is a core requisite to making causal inferences, as the reliability of a finding is increased through corroboration (Maniadis et al., 2014). It is only "*when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated*" (Ioannidis, 2008, p. 640) or as Hume phrases it *"When one particular species of event has always, in all instances, been conjoined with another... We then call the one Object, Cause; the other Effect"* (Hume, 1975, pp. 74-75). Replication is the recreation of empirical evidence and thereby a mitigation of the issues with uncertainty and variation and it is what is required in order to create

generalisable rules for practice to adhere to (Dyckman & Zeff, 2014). Without replications we risk facing a dilemma where *"the cognitive biases of individuals, combined with biases inherent in the review process and the academy in general, foster an environment where the placement of the first research bricks affect the whole wall"* (Bamber et al., 2000, p. 103). As such, we argue that the presumption of causality in contemporary performance measurement has moved prematurely into a stage of normal science as there was not enough empirical evidence for a paradigm to take form.

6.2 Qualitative research on causality

As such, we find support for the argument by Vaivio (2008) that 'theories', or empirical findings in management accounting research is limited in time, space and population, or as Vaivio (2008, p. 69) frames it *"It is not supposed to be a universally valid construct... And it is not supposed to be an eternal construction that stands firm against the ravages of time. Instead, theories are born, have a lifespan and die"*. Questions as those raised by Ittner (2014) might therefore be impossible to provide a 'true' causal answer to.

If Vaivio (2008) and also Nørreklit (2000) are right, it is not necessarily relevant to ask if BSC causally improves performance, but instead to uncover *when* and *how* a BSC improves organisational performance. From this perspective, the BSC has a potential under certain circumstances to improve financial performance, but it is not certain. Which it would have been if the association was claimed to be causal.

However, not all research approaches fit all research questions and in this case, it can be questioned if the hypothetico-deductive method of PMAR is suitable to answer the *when* and *how* a BSC or the use of non-financial performance measures improves financial performance. It might be that a qualitative or mixed method approach could provide more interesting and relevant answers. PMAR and the hypothetico-deductive method requires a stable environment if generalisations are to hold, and concepts need to have a precise and clear conceptual definition. This is a situation that is not always the case. For example, the BSC exists in many forms and variations, which render it difficult to attaint construct validity (Speckbacher, Bischof, & Pfeiffer, 2003).

It can be argued that the value of contemporary performance measurement is dependent on the actors and the organisation implementing it (Mitchell, Nørreklit, & Raffnsøe-Møller, 2016; Nørreklit, 2017), which implies that the success of a PMS, in terms of improving financial performance, cannot be an universal law. Qualitative research or a mixed method approach might therefore be a more appropriate method for studying the relationship between contemporary performance measurement and financial performance, or, specifically how the use of non-financial measures is operationalised in a successful manner. In contrast to NHST, qualitative research accepts and can deal with variation and uncertainty, as it perceives 'theory' as a local

description and explanation as well as a temporal creation instead of an eternal construct (Vaivio, 2008).

The findings of this paper are in accordance with the claim by Mitchell et al. (2016) and Nørreklit, Nørreklit, and Mitchell (2016) that the success of PMSs is not externally given due to causality and that it is unlikely that the functioning of the local practice can be improved by incorporating the answers to causal questions. It is instead something that is created in the local practices by human actors through a system of operating generalisations in the establishment or construction of local causalities (i.e. construct causality (Nørreklit, 2017; Nørreklit et al., 2012)) and it is these operating generalisations that we need to study and understand, as this will help researchers in conceptualising when, how and why PMSs *sometime* improve financial performance. We believe that qualitative and mixed method research is suitable for investigating practice from this perspective and this is why qualitative research remains a cornerstone for developing, improving and producing relevant knowledge for practice (Malina, Nørreklit, & Selto, 2011; Nørreklit, 2014; Vaivio, 2008). It might also provide a deeper insight and explanation to why practitioners appear unwilling to rely on causal relations between non-financial and financial measures (Albertsen & Lueg, 2014; Gates, 1999; Greiling, 2010; Ittner et al., 2003).

7. Conclusion

What is considered to be normal science within the disciplinary matrix of contemporary performance measurement research appears to overlook the inconvenient evidence of inconsistency and uncertainty regarding the existence of causality between non-financial measurement and financial performance. The analysis evidence an unorganised and unstructured body of empirical evidence on cause-and-effect relations providing divergent results on the existence of causality. The causal question of whether non-financial measures are a driving force in financial performance remains unanswered. Yet causality remains a cornerstone in contemporary performance measurement theory. However, without consistent and clear empirical evidence on cause-and-effect relations between non-financial measures and financial performance, we must question the claim of contemporary performance measurement being proactive of the future. It is as such unlikely that contemporary performance measurement can predict and control future financial performance by controlling for the causes i.e. non-financial measures, which was the ultimate ambition of contemporary PMS.

Appendix A

An overview of selected journals

Journal acronyms	Journal name	Journal ranking by		ABI/Inform	Business Souce Complete
		ABDC (2013)	ABS (2010)		
MAR	Management Accounting Research	A*	3	1992 - 2016	
AOS	Accounting, Organizations and Society	A*	4	1976 - 2016	
BAR	British Accounting Review	A	3	1991 - 2016	
AAA	Accounting Auditing and Accountability Journal	A	3	1992 - 2016 (1year delay)	
JMAR	Journal of Management Accounting Research	A	2		1989 - 2016
TAR	The Accounting Review	A*	4		1926 - 2016
JAR	Journal of Accounting Research	A*	4	1972 - 2016	
CAR	Contemporary Accounting Research	A*	3		1984 - 2016 (6month delay)
	ABACUS	A	3	1973 - 2016	
EAR	European Accounting Review	A*	3		1992 - 2016
RAS	Review of Accounting Studies	A*	4		1996 - 2016 (12month delay)

Appendix B

A list of the articles in the sample

- Banker, R. D., & Mashruwala, R. (2007). The Moderating Role of Competition in the Relationship between Nonfinancial Measures and Future Financial Performance*. *Contemporary Accounting Research*, 24(3), 763-793.
- Banker, R. D., Potter, G., & Srinivasan, D. (2000). An empirical investigation of an incentive plan that includes nonfinancial performance measures. *The Accounting Review*, 75(1), 65-92.
- Davis, S., & Albright, T. (2004). An investigation of the effect of balanced scorecard implementation on financial performance. *Management Accounting Research*, 15(2), 135-153.
- Dikolli, S. S., Kinney, W. R., & Sedatole, K. L. (2007). Measuring Customer Relationship Value: The Role of Switching Cost*. *Contemporary Accounting Research*, 24(1), 93-132.
- Dikolli, S. S., & Sedatole, K. L. (2007). Improvements in the information content of nonfinancial forward-looking performance measures: a taxonomy and empirical application. *Journal of Management Accounting Research*, 19(1), 71-104.
- Hoque, Z. (2005). Linking environmental uncertainty to non-financial performance measures and performance: a research note. The British Accounting Review, 37(4), 471-481.
- Ittner, C. D., & Larcker, D. F. (1998). Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction. *Journal of Accounting Research*, 1-35.
- Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, organizations and society*, 28(7), 715-741.
- Malina, M. A., Nørreklit, H., & Selto, F. H. (2007). Relations among Measures, Climate of Control, and Performance Measurement Models*. *Contemporary Accounting Research*, 24(3), 935-982.
- Nagar, V., & Rajan, M. V. (2001). The revenue implications of financial and operational measures of product quality. *The Accounting Review*, 76(4), 495-513.
- Perera, S., Harrison, G., & Poole, M. (1997). Customer-focused manufacturing strategy and the use of operations-based non-financial performance measures: a research note. *Accounting, organizations and society*, 22(6), 557-572.
- Said, A. A., HassabElnaby, H. R., & Wier, B. (2003). An empirical investigation of the performance consequences of nonfinancial measures. *Journal of Management Accounting Research*, 15(1), 193-223.

- Sedatole, K. L. (2003). The effect of measurement alternatives on a nonfinancial quality measure's forward-looking properties. *The Accounting Review*, 78(2), 555-580.
- Smith, R. E., & Wright, W. F. (2004). Determinants of customer loyalty and financial performance. *Journal of Management Accounting Research*, 16(1), 183-205.
- Wiersma, E. (2008). An exploratory study of relative and incremental information content of two non-financial performance measures: Field study evidence on absence frequency and on-time delivery. *Accounting, organizations and society*, 33(2), 249-265.

References

Albertsen, O. A., & Lueg, R. (2014). The Balanced Scorecard's missing link to compensation: A literature review and an agenda for future research. *Journal of Accounting & Organizational Change, 10*(4).

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering'metrics: the path from cause to effect*. New Jersey: Princeton University Press.

Anscombe, G. E. M. (1958). On brute facts. *Analysis*, 69-72.

Atkinson, H. (2006). Strategy implementation: a role for the balanced scorecard? *Management Decision, 44*(10), 1441-1460.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454.

Balakrishnan, R., & Penno, M. (2014). Causality in the context of analytical models and numerical experiments. *Accounting, organizations and society, 39*(7), 531-534.

Baldvinsdottir, G., Mitchell, F., & Nørreklit, H. (2010). Issues in the relationship between theory and practice in management accounting. *Management Accounting Research, 21*(2), 79-82.

Bamber, L. S., Christensen, T. E., & Gaver, K. M. (2000). Do we really 'know' what we think we know? A case study of seminal research and its subsequent overgeneralization. *Accounting, organizations and society, 25*(2), 103-129.

Banker, R. D., Potter, G., & Srinivasan, D. (2000). An empirical investigation of an incentive plan that includes nonfinancial performance measures. *The Accounting Review, 75*(1), 65-92.

Barnabè, F., & Busco, C. (2012). The causal relationships between performance drivers and outcomes: Reinforcing balanced scorecards' implementation through system dynamics models. *Journal of Accounting & Organizational Change, 8*(4), 528-538.

Bird, A. (2013). Thomas Kuhn. In N. Z. Edward (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 ed.).

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Chan, T. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered reports at AIMS neuroscience and beyond. *AIMS Neuroscience, 1*(1), 4-17.

Chawla, D. S. (2017, 19 September 2017). 'One-size-fits-all' threshold for P values under fire. *Nature*. Retrieved from https://www.nature.com/news/one-size-fits-all-threshold-for-p-values-under-fire-1.22625

Chenhall, R. H., & Smith, D. (2011). A review of Australian management accounting research: 1980–2009. *Accounting & Finance, 51*(1), 173-206.

Chua, W. F. (1986). Radical developments in accounting thought. *Accounting review*, 601-632.

Crosby, L., & Sheery, L. (2006). Cause and effect. *Marketing Management, 15*(3), 12-13.

Davis, S., & Albright, T. (2004). An investigation of the effect of balanced scorecard implementation on financial performance. *Management Accounting Research, 15*(2), 135-153.

De Haas, M., & Kleingeld, A. (1999). Multilevel design of performance measurement systems: enhancing strategic dialogue throughout the organization. *Management Accounting Research, 10*(3), 233-261.
Dyckman, T. R., & Zeff, S. A. (2014). Some methodological deficiencies in empirical research articles in accounting. *Accounting horizons, 28*(3), 695-712.
Edwards, P. (1972). The encyclopaedia of philosophy (Vols 1–8). US: Macmillian Publishing Co., Inc. & The Free Press.
Evans, J. H., Feng, M., Hoffman, V. B., Moser, D. V., & Stede, W. A. (2015). Points to Consider When Self-Assessing Your Empirical Accounting Research. *Contemporary Accounting Research, 32*(3), 1162-1192.
Fahrbach, L. (2005). Understanding Brute Facts. *An International Journal for Epistemology, Methodology and Philosophy of Science, 145*(3), 449-466. doi:10.1007/s11229-005-6200-7
Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.
Franco-Santos, M., Lucianetti, L., & Bourne, M. (2012). Contemporary performance measurement systems: A review of their consequences and a framework for research. *Management Accounting Research, 23*(2), 79-119.
Friedman, M. (1953). The methodology of positive economics. *Essays in positive economics, 3*(3).
Gassen, J. (2014). Causal inference in empirical archival financial accounting research. *Accounting, organizations and society, 39*(7), 535-544.
Gates, S. (1999). *Aligning strategic performance measures and results*. New York: The Conference Board
Gelman, A. (2011). Causality and Statistical Learning. *American Journal of Sociology, 117*(3), 955-966.
Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research a Bayesian perspective. *Journal of Management, 41*(2), 632-643.
Gelman, A. (2015). Statistics and the crisis of scientific replication. *Significance, 12*(3), 39-41.
Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460.
Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics, 15*(3), 373-390.
Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.
Gigerenzer, G., & Marewski, J. N. (2015). Surrogate Science The Idol of a Universal Method for Scientific Inference. *Journal of Management, 41*(2), 421-440.
Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine, 8*(341), 341ps312-341ps312.
Greiling, D. (2010). Balanced Scorecard implementation in German non-profit organisations. *International Journal of Productivity and Performance Management, 59*(6), 534-554.
Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods, 12*(3), 179-185.
Hoque, Z. (2014). 20 years of studies on the balanced scorecard: Trends, accomplishments, gaps and opportunities for future research. *The British Accounting Review, 46*(1), 33-59.
Hubbard, R., & Lindsay, R. M. (2013). The significant difference paradigm promotes bad science. *Journal of Business Research, 66*(9), 1393-1397.
Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals* (3. ed., repr. / with text rev. and notes by P.H. Nidditch ed.). Oxford: Clarendon.
Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.
Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648.
Ittner, C. D. (2014). Strengthening causal inferences in positivist field studies. *Accounting, organizations and society, 39*(7), 545-549.

Ittner, C. D., & Larcker, D. F. (1998). Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction. *Journal of accounting research*, 1-35.
Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, organizations and society, 28*(7), 715-741.
Janeš, A. (2014). Empirical verification of the balanced scorecard. *Industrial Management & Data Systems, 114*(2), 203-219.
Kaldor, N. (1961). Capital accumulation and economic growth. In D. C. Hague (Ed.), *The theory of capital* (pp. 177-222): Springer.
Kane, E. J. (1984). Why journal editors should encourage the replication of applied econometric research. *Quarterly Journal of Business and Economics, 23*(1), 3-8.
Kaplan, R. S., & Norton, D. P. (1992). The balanced scorecard--measures that drive performance. *70*(1), 71.
Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: translating strategy into action*. Boston, Mass: Harvard Business Press.
Kasperskaya, Y., & Tayles, M. (2013). The role of causal links in performance measurement models. *Managerial Auditing Journal, 28*(5), 426-443.
Kuhn, T. S. (1970). *The structure of scientific revolutions* (2 ed.). Chicago: University of Chicago Press.
Lachmann, M., Trapp, I., & Trapp, R. (2017). Diversity and validity in positivist management accounting research—A longitudinal perspective over four decades. *Management Accounting Research*.
Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research, 11*(1), 33.
Lueg, R., & Nørreklit, H. (2013). Performance measurement systems - beyond generic actions. In F. Mitchell, H. Nørreklit, & M. Jakobsen (Eds.), *The Routledge Companion to Cost Management* (pp. 342-359). Abingdon, Oxon: Routledge.
Luft, J. (2009). Nonfinancial information and accounting: A reconsideration of benefits and challenges. *Accounting horizons, 23*(3), 307-325.
Luft, J., & Shields, M. D. (2014). Subjectivity in developing and validating causal explanations in positivist accounting research. *Accounting, organizations and society, 39*(7), 550-558.
Lukka, K. (2014). Exploring the possibilities for causal explanation in interpretive research. *Accounting, organizations and society*.
Malina, M. A., Nørreklit, H., & Selto, F. H. (2007). Relations among Measures, Climate of Control, and Performance Measurement Models. *Contemporary Accounting Research, 24*(3), 935-982.
Malina, M. A., Nørreklit, H. S., & Selto, F. H. (2011). Lessons learned: advantages and disadvantages of mixed method research. *Qualitative Research in Accounting & Management, 8*(1), 59-71.
Malmi, T. (2001). Balanced scorecards in Finnish companies: a research note. *Management Accounting Research, 12*(2), 207-220.
Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review, 104*(1), 277-290.
McNutt, M. (2014). Journals unite of reproducibility. *Science, 346*(6210), 679.
Micheli, P., & Mari, L. (2014). The theory and practice of performance measurement. *Management Accounting Research, 25*(2), 147-156.
Mitchell, F., Nørreklit, H., & Raffnsøe-Møller, M. (2016). A pragmatic constructivist approach to accounting practice and research. *Qualitative Research in Accounting & Management, 13*(3).
Nuzzo, R. (2014). Statistical errors. *Nature, 506*(7487), 150-152.
Nørreklit, H. (2000). The balance on the balanced scorecard a critical analysis of some of its assumptions. *Management Accounting Research, 11*(1), 65-88.
Nørreklit, H. (2014). Quality in qualitative management accounting research. *Qualitative Research in Accounting & Management, 11*(1), 29-39.

Nørreklit, H. (2017). *A Philosophy of Management Accounting: A Pragmatic Constructivist Approach*. New York: Routledge.
Nørreklit, H., Nørreklit, L., & Mitchell, F. (2010). Towards a paradigmatic foundation for accounting practice. *Accounting, Auditing & Accountability Journal, 23*(6), 733-758.
Nørreklit, H., Nørreklit, L., & Mitchell, F. (2016). Understanding practice generalisation–opening the research/practice gap. *Qualitative Research in Accounting & Management, 13*(3).
Nørreklit, H., Nørreklit, L., Mitchell, F., & Bjørnenak, T. (2012). The rise of the balanced scorecard! Relevance regained? *Journal of Accounting & Organizational Change, 8*(4), 490-510.
Otley, D. (1999). Performance management: a framework for management control systems research. *Management Accounting Research, 10*(4), 363-382.
Papaioannou, D., Sutton, A., Carroll, C., Booth, A., & Wong, R. (2010). Literature searching for social science systematic reviews: consideration of a range of search techniques. *Health Information & Libraries Journal, 27*(2), 114-122.
Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics, 6*(2), 1-59.
Perera, S., Harrison, G., & Poole, M. (1997). Customer-focused manufacturing strategy and the use of operations-based non-financial performance measures: a research note. *Accounting, organizations and society, 22*(6), 557-572.
Said, A. A., Hassabelnaby, H. R., & Wier, B. (2003). An empirical investigation of the performance consequences of nonfinancial measures. *Journal of Management Accounting Research, 15*(1), 193-223.
Scapens, R. W., & Bromwich, M. (2001). Editorial Report—Management Accounting Research: the first decade. *Management Accounting Research, 12*(2), 245-254.
Shields, M. D. (1997a). Research in management accounting by North Americans in the 1990s. *Journal of Management Accounting Research*(9), 3-61.
Shields, M. D. (1997b). Research in management accounting by North Americans in the 1990s. *Journal of Management Accounting Research, 9*, 3.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366.
Simon, J. L. (1970). The Concept of Causality in Economics. *Kyklos, 23*(2), 226-254.
Slife, B. D., & Williams, R. N. (1995). *What's behind the research?: Discovering hidden assumptions in the behavioral sciences*. London: Sage.
Speckbacher, G., Bischof, J., & Pfeiffer, T. (2003). A descriptive analysis on the implementation of balanced scorecards in German-speaking countries. *Management Accounting Research, 14*(4), 361-388.
Vaivio, J. (2008). Qualitative management accounting research: rationale, pitfalls and potential. *Qualitative Research in Accounting & Management, 5*(1), 64-86.
van der Meer-Kooistra, J., & Vosselman, E. (2012). Research paradigms, theoretical pluralism and the practical relevance of management accounting knowledge. *Qualitative Research in Accounting & Management, 9*(3), 245-264.
Van der Stede, W. A. (2014). A manipulationist view of causality in cross-sectional survey research. *Accounting, organizations and society, 39*(7), 567-574.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-values: context process, and purpose. *The American Statistician*.
Wiersma, E. (2008). An exploratory study of relative and incremental information content of two non-financial performance measures: Field study evidence on absence frequency and on-time delivery. *Accounting, organizations and society, 33*(2), 249-265.
Wold, H. (1954). Causality and econometrics. *Econometrica: journal of the Econometric Society*, 162-177.
Wright, G. H. v. (1994). *Myten om fremskridtet: tanker 1987-92 med en intellektuel selvbiografi* (2. opl. ed.). Kbh.: Munksgaard.

Zimmerman, J. L. (2001). Conjectures regarding empirical managerial accounting research. *Journal of Accounting and Economics, 32*(1), 411-427.

Chapter 3

IS THE VALIDITY OF POSITIVISTIC MANAGEMENT ACCOUNTING RESEARCH EXPOSED TO QUESTIONABLE RESEARCH PRACTICES?

Author: Kristian Mohr Røge

Abstract A recent paper in Management Accounting Research (MAR) claimed that the validity of positivistic management accounting research (PMAR) has increased significantly during the last four decades.

We argue that this is a misrepresentation of reality as the current crisis of irreproducible statistical findings is not addressed. The reliability and validity of statistical findings are under an increasing pressure due to the phenomenon of Questionable Research Practices (QRPs). It is a phenomenon argued to increase the ratio of false-positives through a distortion of the hypothetico-deductive method in favour of a researcher's own hypothesis. This phenomenon is known to be widespread in the social sciences. We therefore conduct a meta-analysis on susceptibility of QRPs on the publication practices of PMAR, and our findings give rise to reasons for concern as there are indications of a publication practice that (unintentionally) incentivises the use of QRPs. It is therefore rational to assume that the ratio of false-positives is well-above the conventional five-per cent ratio. To break the bad equilibrium of QRPs, we suggest three different solutions and discuss their practical viability.

Keywords: Philosophy of science; Statistical methods; Questionable Research Practices (QRPs); Hypothetico-deductive method; NHST

1. Introduction

Statistical inferences independent of scientific field are proving to be fragile, unreliable, lacking replicability and facing emerging problems with generalising 'lab' findings to a real-world setting[8] (Gelman & Loken, 2014b; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). The irreproducibility of statistical findings has been found to be widespread (e.g., strategic management (Bergh, Sharp, Aguinis, & Li, 2017), economics (Camerer et al., 2016), general management (Banks, O'Boyle, et al., 2016; Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016), medicine (Begley & Ellis, 2012; Ioannidis, 2005; Prinz, Schlange, & Asadullah, 2011), neuroscience (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014)). The irreproducibility of statistical research is argued to be caused by Questionable Research Practices (QRPs), which is a phenomenon claimed to distort the hypothetico-deductive method in favour of a researcher's own hypothesis with the side effect of increasing the probability of a false-positive (Chambers et al., 2014; Ioannidis, 2005; Simmons et al., 2011).

Studies have confirmed the QRPs activities among business school researchers (Butler, Delaney, & Spoelstra, 2017; O'Boyle, Banks, & Gonzalez-Mulé, 2017) arguing that playing with numbers, playing with models and playing with hypotheses are not uncommon. Scholars have explained the existence of QRPs in three ways: the inadequate training of researchers, the pressure and incentives to publish in certain outlets, and the demand and expectations of journal editors and reviewers.

As the *sine qua non* method of positivistic management accounting research (PMAR) is null-hypothesis testing (NHST) (Chua, 1986; Lachmann, Trapp, & Trapp, 2017; Lindsay, 1994; Merchant, 2010; Shields, 1997; Van der Stede, Young, & Chen, 2005), it implies that QRPs is a potential threat to the validity of causal claims that PMAR intends to make (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2014). It would be naïve to assume that researchers within PMAR are somehow resilient to QRPs if institutional actors, such as academic journals and publication counting deans, (unintentionally) incentivise QRP activities.

Unfortunately, the very attempt to uncover a 'true' rate of QRPs in published research is obscured by the very practices that make them questionable, as they tend not to be conducted transparently due to either misreporting or lack of reporting (Banks, O'Boyle, et al., 2016). But, if the publication practice of academic journals allows QRPs, then we must assume that QRPs have already invaded that particular scientific field.

[8] The October 19, 2013 issue of *The Economist* 'Unreliable research: Trouble at the lab' provides a lengthy critique of the inability to replicate scientific research. While ScienceNews wrote: It's Science's dirtiest secret: the "scientific method" of testing hypotheses by statistical analysis stands on a flimsy foundation" (Siegfried, 2010). Furthermore, a study by Camerer et al. (2016) published in Science tried to replicate 18 studies that were published in the *American Economic Review* and the *Quarterly Journal of Economics*. They *"found a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original..."* (Camerer et al., 2016, p. 1433).

In this light, we find it reasonable and necessary to investigate *if* the publication practices of high-ranking journals in the field of management accounting allow QRPs or whether they somehow have managed to be resilient to QRPs. As these high-ranking journals represent the ideal of scientific integrity and as the editors and reviewers are gatekeepers to scientific integrity, it implies that the solution to QRPs also lies here. QRPs are a paradox at work, because to live up to the positivistic image of 'pure science', academic journals and researchers may find themselves transgressing this very ideal (Butler et al., 2017).

The purpose of this paper is therefore to shed light on the likelihood of QRPs within PMAR and, in doing so, we try to answer the following research question: How susceptible are the publication practices of PMAR to the phenomenon of QRPs? To develop an analytical framework that can provide an answer to this question, we draw on the accumulated knowledge on QRPs from a wide range of scientific fields (e.g. Banks, O'Boyle, et al., 2016; Garud, 2015; Gelman & Loken, 2014a; Gigerenzer & Marewski, 2015; Ioannidis, 2005; Kerr, 1998; Nuzzo, 2014; Simmons et al., 2011).

The next section provides an extensive clarification of the phenomenon of QRPs and their consequences for the hypothetico-deductive method. The third section outlines the analytical framework developed oi section two and clarifies the data selection. Section four presents and discusses the findings, while section five concludes on the analysis and suggests how to mitigate a further development of QRPs in PMAR.

2. Questionable Research Practices – what is that?

The scientific system, or the 'publish or perish' culture, is claimed to have fostered a range of QRPs within hypothetico-deductive method (Chambers et al., 2014; Leung, 2011; Lindsay, 1994) and, as a result, the research community has expressed two main concerns. The first concern is a research bias: *"the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced"* (Ioannidis, 2005, p. 41). The second concern addresses the 'publish or perish' thought in research, where *"QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage"* (John, Loewenstein, & Prelec, 2012, p. 524). Unfortunately, it appears that QRPs are not limited to a small subsection of the scientific community but are, in fact, widespread and by some considered 'defensible' (Butler et al., 2017; John et al., 2012; Martinson, Anderson, & de Vries, 2005; Starbuck, 2016).

By convention, researchers are ultimately perceived to be 'pure', that is, individuals motivated solely by the acquisition of knowledge. However, this utopian idea of a researcher is in reality flawed as researchers have human needs, desires and motives, just like non-researchers (Mahoney, 1977). Furthermore, the scientific ecosystem, in which researchers work, is built and

maintained by a number of stakeholders (i.e. universities, funding bodies, industry stakeholders and publishers) whose interest may diverge from pure knowledge accumulation (Hardwicke, Jameel, Jones, Walczak, & Weinberg, 2014). Unfortunately, it appears that the incentives embedded in the scientific system do not adequately account for these human factors and, thus, reward individuals that are lucky or willing to 'play the game' of QRPs (Chambers et al., 2014; Hardwicke et al., 2014; Nosek, Spies, & Motyl, 2012; Starbuck, 2016).

QRPs are defined as a range of activities that, intentionally or unintentionally, distorts the hypothetico-deductive method in favour of a researcher's own hypothesis (Chambers et al., 2014; Hardwicke et al., 2014; John et al., 2012). These practices are typically *P-hacking, low statistical power, Hypothesising After the Results are Known (HARKing), publication bias, a lack of data sharing and lack of replications*, and they are claimed to increase the probability of making a false-positive finding (Ioannidis, 2005; Maniadis, Tufano, & List, 2014; Simmons et al., 2011). If QRPs are employed on a large scale, it would have a devastating impact on the validity of an entire field of scientific inquiry.

In the next section, we explore each QRP in a greater detail.

2.1 P-hacking and low statistical power: A search for significance

"If you torture the data long enough, it will confess."

Ronald Coase (Tullock, 2001, p. 205)

In research, NHST has become equated with scientific rigour and perceived as *the* touchstone for establishing knowledge. It is considered the *sine qua non* of the 'scientific method' for making scientific inferences (Gigerenzer & Marewski, 2015; Hubbard & Lindsay, 2013; Lindsay, 1994).

To clarify, NHST is a method where a researcher seeks to reject a straw-man null hypothesis as evidence in favour of some favoured alternative hypothesis based on the obtained *P* value. The *American Statistical Association* informally describes *P* value as the probability of a statistical summary of the data being equal to or more extreme than its observed value under a specified statistical model (Wasserstein & Lazar, 2016). A *P* value therefore measures the incompatibility between a set of data and a proposed model for the data. That is, the smaller the *P* value, the greater the statistical incompatibility of the data with the null hypothesis – if the underlying assumptions used to calculate the *P* value hold. However, *P* values do not measure the probability for the studied hypotheses to be true, nor the probability for the data to be produced by random chance alone (Wasserstein & Lazar, 2016). Instead this depends on a range of parameters, for instance, the prior probability of it being true (before doing the study), the statistical power of the study and the level of significance (Ioannidis, 2005; Nuzzo, 2014).

The luring danger to NHST is false-positives (Simmons et al., 2011; Simonsohn, Nelson, & Simmons, 2014). Replication of studies would ideally protect science against false-positives as intuitively a result should only be trusted if it is corroborated by many different studies, such as

"when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated" (Ioannidis, 2008, p. 640). The intuition of this argument is the premise that false-positive findings tend to require many failed attempts at the prevailing rule of a significance level of 0.05. From this perspective, a researcher studying a non-existent effect would on average observe a false-positive finding only once in 20 studies (Simonsohn et al., 2014), and corroborated studies would therefore in fact appear to be true.

However, the seminal work by Ioannidis (2005) in *PLoS Medicine* questioned this premise, an argument which since has been reclaimed in *American Economic Review* (Maniadis et al., 2014). Ioannidis presented a Bayesian argument for why the irreproducibility crisis in science should not come as a surprise, as the false discovery rate (FDR) is likely to be much higher than the assumed ratio of five percent. This claim has received enormous attention from the broad scientific community manifested by over 4,900 Google Scholar citations[9]. The theorem is constructed around the two terms: positive predictive value (PPV) and FDR:

$$PPV = 1 - FDR = \frac{(1-\beta)R + \mu\beta R}{(1-\beta)R + \mu\beta R + \alpha + \mu(1-\alpha)}$$

The PPV is the probability that a finding is *true*, while the FDR is the probability that a finding is *false*. The theorem informs us that as either type I error rate (α) increases, the statistical power (1 – β) decreases or the ratio of false to true hypotheses (R) decreases, the FDR will increase. In addition, Ioannidis modelled all sources of bias into a single factor μ, which is the proportion of null hypotheses that would not have been claimed as discoveries in the absence of bias, but which ended up as such, because of it. Thus, if μ increases, the PPV would go down and the FDR would go up. Therefore, with an increasing bias, the chance of a research finding being true diminishes. Table 1 illustrates different calculations of PPV as a function of R and for various settings of β, α and μ to evidence the significance of the ratio of false to true hypotheses (R) and the level of bias (μ), as even with a very low alpha and high statistical power (1 – β), it proves difficult to attain a satisfactory PPV:

[9] Google Scholar – 31/08-2017.

Table 1. The PPV estimate as a function of prior R and for various settings of β, α and μ

	Power 80%, Alpha = 0.01				**Power 60%, Alpha = 0.01**			
	u = 0.05	*u = 0.1*	*u = 0.3*	*u = 0.5*	*u = 0.05*	*u = 0.1*	*u = 0.3*	*u = 0.5*
R	PPV				PPV			
0.1	0.58	0.43	0.22	0.15	0.51	0.37	0.19	0.14
0.2	0.73	0.60	0.36	0.26	0.68	0.54	0.32	0.24
0.3	0.80	0.69	0.46	0.35	0.76	0.64	0.41	0.32
0.4	0.84	0.75	0.53	0.42	0.81	0.70	0.48	0.39
0.5	0.87	0.79	0.58	0.47	0.84	0.75	0.54	0.44
0.6	0.89	0.82	0.63	0.52	0.86	0.78	0.58	0.49
0.7	0.91	0.84	0.66	0.56	0.88	0.80	0.62	0.53
0.8	0.92	0.86	0.69	0.59	0.89	0.82	0.65	0.56
0.9	0.92	0.87	0.72	0.62	0.90	0.84	0.68	0.59

	Power 80%, Alpha = 0.05				**Power 60%, Alpha = 0,05**			
	u = 0.05	*u = 0.1*	*u = 0.3*	*u = 0.5*	*u = 0.05*	*u = 0.1*	*u = 0.3*	*u = 0.5*
R	PPV				PPV			
0.1	0.45	0.36	0.20	0.15	0.39	0.31	0.18	0.13
0.2	0.62	0.53	0.34	0.26	0.56	0.47	0.30	0.23
0.3	0.71	0.63	0.44	0.34	0.66	0.57	0.39	0.31
0.4	0.77	0.69	0.51	0.41	0.72	0.64	0.46	0.38
0.5	0.81	0.74	0.56	0.46	0.76	0.69	0.52	0.43
0.6	0.83	0.77	0.61	0.51	0.79	0.73	0.56	0.48
0.7	0.85	0.80	0.64	0.55	0.82	0.76	0.60	0.52
0.8	0.87	0.82	0.67	0.58	0.84	0.78	0.63	0.55
0.9	0.88	0.84	0.70	0.61	0.85	0.80	0.66	0.58

The argument by Ioannidis (2005) has been further substantiated by evidence produced by simulations demonstrating that, in a singly study, only a few changes in data analysis decisions could increase the FDR to approximately 60 percent (Simmons et al., 2011). The study shows that *P*-hacking activities upends the assumption about the number of failed studies that is required to produce a false-positive finding. Furthermore, Simmons et al. (2011) argued that *P*-hacking, in the form of undisclosed data flexibility in collection and analysis, allowed them to present almost any hypothesis as significant.

P-hacking should be understood as a term that is confined by a researcher's collective actions taken to exploit ambiguity in a pursuit for statistical significance (Halsey, Curran-Everett, Vowler, & Drummond, 2015; Motulsky, 2015; Nuzzo, 2014; Simmons et al., 2011); in other words, *P*-hacking is *any* and *all* post-hoc changes to the data analysis. Figure 1 demonstrates the process of *P*-hacking.

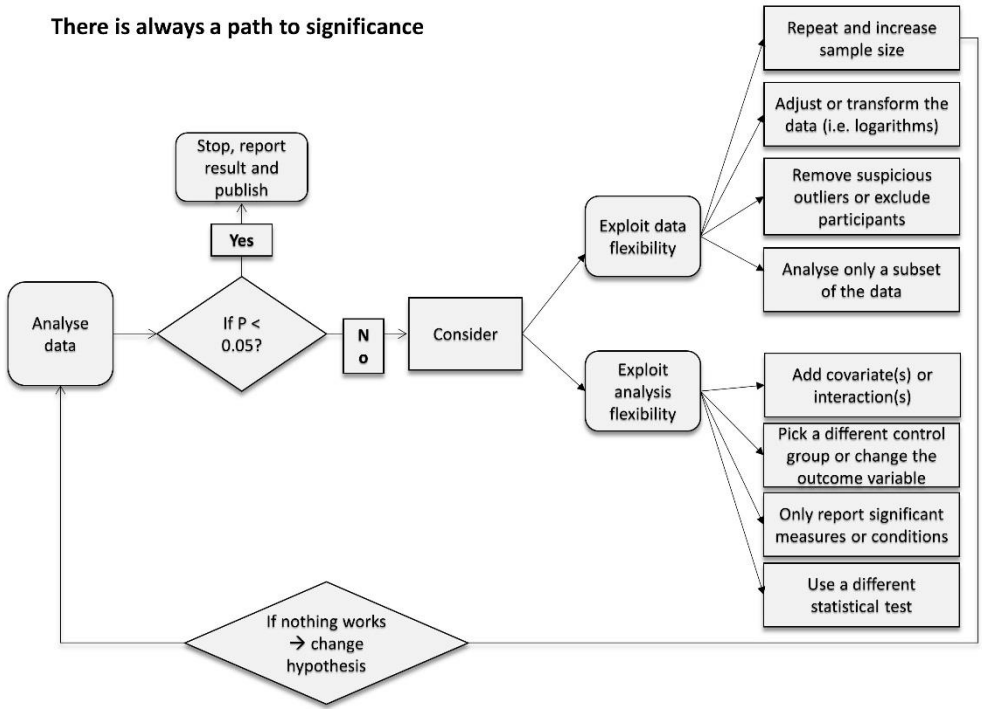


Fig. 1. What counts as P-hacking? Any and all post-hoc changes, as with enough flexibility, there will always be a path leading to significant effects (Motulsky, 2015; Simmons et al., 2011)

P-hacking seems to have emerged as a long-term consequence in a world of publication-counting universities along with a positive publication bias in academic outlets. These two factors provide near perfect incentives for researchers to maximise publications by chasing their data for significance. To protect science against *P*-hacking, it has been suggested that transparency be increased in terms of providing access to raw data and data analysis (Nuzzo, 2014); the *American Statistical Association* further argues that proper inferences require *full* reporting and transparency:

> *"P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values... renders the reported p-values essentially uninterpretable.... data dredging, significance chasing, significance questing, selective inference and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided... Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting."*
>
> (Wasserstein & Lazar, 2016, pp. 9-10)

2.2 HARKing: 'Hypothesising After the Results are Known'

"A reader quick, keen and leery
Did wonder, ponder and query
When results clean and tight
Fit predictions just right
If the data preceded the theory"

Anonymous (Kerr, 1998, p. 196)

In psychology research, HARKing has been a topic of debate for quite some time (Kerr, 1998). HARKing involves generating a hypothesis from the dataset, by uncovering an intriguing significant relationship, typically through innovative statistical analyses (Motulsky, 2015), and then presenting it as an *a priori* hypothesis (Anonymous, 2015; Chambers et al., 2014; Kerr, 1998). The HARKing debate has since reached general management research, where a provocation and provocateur's piece in *Journal of Management Inquiry* discusses the process and ethics of HARKing. The piece is an anonymous author's reflection on the ethical aftermath of engaging in HARKing[10]. The author recounts how he and his co-authors were able to find intriguing significant results through an innovative analytical approach but were unable to find support for their original *a priori* hypotheses (Anonymous, 2015). After finding these intriguing results, they rewrote the article as if they had formulated these new hypotheses in advance, justifying this by acknowledging that everybody does it (Butler et al., 2017). However, in the end, the anonymous author felt uncomfortable with this, but one of his co-authors was not tenured, and for his sake, the anonymous author felt highly motivated to see the paper published. The paper ended up being published in an A-journal in their field.

When conducting HARKing, the researcher uses the same dataset for generating the hypothesis and testing it (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). But, a hypothesis should not be tested with the same data from which it was derived; this is essential scientific misconduct, as such hypotheses would be no more than empirical findings disguised as hypotheses (Leung, 2011). In blunter terms, Garud (2015, p. 452) frames it as: *"the practice of presenting post hoc hypothesis as a priori can end up compromising what we know. That is, epistemology compromises ontology. Specifically, this practice can increase the possibility of Type I and Type II errors, misrepresent the truth-value of hypotheses that never were".* HARKing allows for fictitious results becoming an immutable truth, which in the long run could contaminate the entire knowledge pool of a scientific field (Garud, 2015). HARKing also counteracts the communication of valuable information about what did not work, which can challenge the development of valid scientific theories (Kerr, 1998).

HARKing appears to be researchers' long-term response to publication pressure along with a positive publication bias where significant outcomes are more likely to be published.

[10] The identity is known to the editors.

(Francis, 2014; Garud, 2015; Ioannidis, 2005; Simmons et al., 2011). Nevertheless, HARKing is still a practise that violates the percepts of basic scientific method but appears to be accepted as normal, and which is practically impossible to verify empirically (Garud, 2015).

2.3 Publication bias: possible misrepresentation of reality

Significance testing of null hypotheses has long been considered a touchstone for establishing knowledge, and a publication bias appears to have emanated from holding this episteme (Hubbard & Lindsay, 2013; Lindsay, 1994; Sterling, 1959). The bias originates from journals that reject manuscripts because they fail in attaining statistical significant results or in removing insignificant findings during the review process (Chambers et al., 2014). Allegedly this situation has prompted a '*file drawer problem'* where researchers' filing cabinets are believed to be chuck-full of insignificant findings (Simonsohn et al., 2014); consequently, failure to report non-supported hypotheses may lead others to continue their efforts to test the very same hypotheses in subsequent research (Bedeian, Taylor, & Miller, 2010; Garud, 2015; Greenwald, 1975). In the end, false-positives may potentially emerge from the continuous efforts in trying to prove the same hypotheses thereby causing misrepresentation of reality, as reality would predominantly consist of significant findings (Ioannidis, 2008)

Statistical significance is not always a prerequisite, nor is it ever sufficient for establishing research findings as scientifically valuable or meaningful (Gigerenzer & Marewski, 2015; Lindsay, 1994; Ziliak, 2016). This is due to statistical significance not being equivalent to scientific, human or economic significance (Wasserstein & Lazar, 2016); the basis for publication should therefore not be statistical significance but rather the scientific significance of a finding. A positive publication bias is therefore, in truth, an undesirable situation for science. Lykken (1968, pp. 158-159) shares this perception: *"The moral of this story is that the finding of statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an experimental report ought to be published".*

The episteme of NHST has created frustration and tension among journals and editors and an outgoing editor of *Journal of Applied Psychology* expressed his dissatisfaction as: *"Perhaps P values are like mosquitos. They have an evolutionary niche somewhere and no amount of scratching, swatting, or spraying will dislodge them...Investigators must learn to argue for the significance of their results without reference to inferential statistics"* (Campbell, 1982, p. 698). Others saw NHST as a decline in statistical thinking and blamed it on Fisher: *"Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe the almost universal reliance on merely refuting the null hypothesis... is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology"* (Meehl, 1978, p. 817). In an effort to oppose the dominance of NHST, the editors of *Basic and Applied Social Psychology* decided to ban *P* values (Wasserstein & Lazar, 2016) and it is not the

first time in history that a scientific journal has tried an approach of banning *P* value. *The New England Journal of Medicine* (in the 1970s), *Epidemiology and the American Journal of Public Health* (in the 1990s) and the *Publication Manual of the American Psychological Association* have all experimented with bans (Ziliak, 2016).

The common thread of these claims and bans is that we, as researchers, must be able to argue for the scientific significance of our findings (Ziliak & McCloskey, 2004a, 2004b), instead of indulging in what is described by Gigerenzer and Marewski (2015) as *mindless statistical inference*. In the end, a systematic omission of insignificant results would impede the advancement of a scientific field as these findings can contribute to science.

2.4 Lack of data sharing: A neglected possibility to increase integrity and credibility

The natural sciences have always acknowledged reproducibility as a cornerstone of scientific practice and a fundamental form of validation. It is therefore alarming for the validity of science when natural scientists recognise that there are issues as regards the reproducibility of published results, an awareness epitomised by a series in *Nature* entitled "Challenges in irreproducible research"[11] and by the "Reproducibility Initiative"[12], a project intended to identify and reward replications (Halsey et al., 2015). In self-correcting science, valuation of replications is essential, and data sharing might therefore be an all-important aspect of research conduct, as it would allow researchers to verify original analyses, conduct novel analyses or carry out meta-analyses that can corroborate the reliability and magnitude of reported effects (Hardwicke et al., 2014; Tenopir et al., 2011).

In an effort to illustrate how data sharing could corroborate research findings, Silberzahn and Uhlmann (2015) recruited 29 research teams and asked them to answer an identical research question with an identical dataset. What was soon to be evidenced was that research teams approach a research question and a dataset in different ways, and interestingly, produce highly varying results. The research teams were asked to investigate whether dark-skinned players were more likely to receive a red card in football than white-skinned ones. Of the 29 research teams, 20 teams found a statistically significant correlation between skin colour and red cards. However, the findings varied enormously in effect sizes, from a slight (non-significant) tendency to a strong (significant) tendency for referees to give more red cards to dark-skinned than white-skinned players. The variation in results indicates that any single team's results are highly influenced by their subjective choices taken in the analysis phase, a result also theoretically evidenced through a simulation study by Simmons et al. (2011).

[11] For more information on the series see http://www.nature.com/content/nature/24974-01.html
[12] For more information on this project see http://validation.scienceexchange.com/#/

The Silberzahn and Uhlmann (2015) study, published in *Nature*, illustrates that data sharing would in fact not only reinforce our research findings through corroboration but also allow self-correcting irreproducible findings. On the other hand, a lack of data sharing would impede the detection of QRPs and prevent detailed meta-analyses with the overall result of impeding the replication of previous findings. If journals begin promoting a culture for data sharing or retention, we could increase transparency and reproducibility of the scientific process (Munafò et al., 2014).

2.5 Lack of replication: The hallmark of science is self-correction

The importance of replication can be learned from an age-old prayer *"Lord, protect us from what we only think we know"* and from an equal ancient hypothesis that God only helps those who help themselves (Kane, 1984). Replication is what ensures credibility and integrity in research, it is about creating rigorous theories or hypotheses by opposing the accumulation and dissemination of false knowledge (Merton, 1942, 1973).

By convention, 'true' or genuine replication requires a process where the exact same finding is re-examined in the exact same way (Moonesinghe, Khoury, & Janssens, 2007). Often, therefore, genuine replication is impossible, and instead corroboration or indirect supporting evidence are more realistic (Goodman, Fanelli, & Ioannidis, 2016). In financial accounting, a replication is defined as: *"redoing the identical study in the same way but for another sample period, or periods"* (Dyckman & Zeff, 2014, p. 698). A corroboration approach might be the only sustainable way of doing replications in the social sciences; however, according to Moonesinghe et al. (2007), a lurking danger is the tiny distance from this type of replications to QRPs, which in the end might contribute to 'pseudo' replications, which are false-findings corroborated by other false-findings.

In the social sciences it is rather rare to see a replication being published, which is typically accredited to two main factors (Dyckman & Zeff, 2014; Goodman et al., 2016). First, the high cost of searching for errors in empirical research; and second, it is notably less rewarding in terms of reputation and ability to publish. According to Kane (1984, p. 3), choosing a task of replication is "*widely regarded as prima facie evidence of intellectual mediocrity, revealing a lack of creativity and perhaps even a bullying spirit*". Status and promotion are only granted to the publication which is published first (Dyckman & Zeff, 2014; Gigerenzer & Marewski, 2015). Nevertheless, reproducibility remains the cornerstone of the hypothetico-deductive method, because if an empirical finding is to be considered a 'fact', other researchers must be able to observe it, thus strengthening the credibility of the 'fact' (Kane, 1984).

Maniadis et al. (2014) provided theoretical evidence for the importance of replications by demonstrating through Bayesian statistics that a few independent replications dramatically increase the chances of an original finding being true, and they therefore claim that replications are the best solution to the current inference problem. On the other hand, a limited number of

replications would force researchers to generalise from a small number of published studies consequently limiting the ability of research to accumulate 'true' knowledge, in particular if the studies did not 'tell the whole story' or proved unreliable (Bamber, Christensen, & Gaver, 2000). If we consider the importance of replications in creating rigorous theories or hypotheses, it should be defined as *the* 'scientific gold standard' (Jasny, Chin, Chong, & Vignieri, 2011).

3. Research Method

3.1 Previous attempts at identifying the prevalence of QRPs in published research

Various attempts have been made at uncovering an approximation of the 'true' ratio of QRPs in published research.

For instance, O'Boyle et al. (2017) found that in management research the ratio of supported to unsupported hypotheses more than doubled from defended dissertation to journal publication. They evidenced that the increase in predictive accuracy was a result of QRPs through the practices of dropping nonsignificant hypotheses, the addition of significant hypotheses, the reversing of predicted direction of hypotheses and alterations to the data. Another example of QRPs is from a survey of over 2,000 academic psychologists at major U.S. universities, which found that almost half admitted having selectively decided to report studies that 'worked' or decided to collect more data after having examined whether the results were significant (John et al., 2012). A study by Jager and Leek (2013) produced a conservative estimate of the ratio of false-positive findings in published studies in five major medical journals. Their meta-study found the false-positive ratio to be 14 percent (Jager & Leek, 2013), however, in a commentary Ioannidis stated that Jager and Leek (2013) fall into the same "false result" category because "their approach is flawed in sampling, calculations and conclusions" (Ioannidis, 2014). Another commentary by two statisticians substantiated Ioannidis's claims, arguing that the false discovery rate is probably closer to 30 or even 50 percent when adjusting for the sample (Benjamini & Hechtlinger, 2014). As a final example of QRPs, a recent survey of 1,500 researchers, initiated by Nature, investigated the current reproducibility crisis in the natural sciences (Baker, 2016). They found that more than 70 percent of researchers have failed in reproducing another researcher's experiment, and more than half had failed in reproducing their own results. It was therefore not surprising, when 52 percent argued for a significant replication crisis, while 38 percent saw it as only a slight crisis, and 7 percent argued for the non-existence of a crisis (3 percent did not know). When asked which factors contributed to irreproducible research, the main contributing factors argued were selective reporting, pressure to publish, low statistical power and/or poor analysis, etc.; and when asked which factors could boost reproducibility, a better understanding of statistics was emphasised by more than 80 percent of the respondents.

These studies illustrate that in many scientific fields QRPs are a reality, but the exact extent of this QRPs is unknown.

3.2 What are the characteristics of a publication practice that allows the existence of QRPs?

Based on section two, we can identify a set of characteristics that are necessary conditions in a publication system for QRPs to take root, and if the publication system of PMAR exhibits such characteristics, we can expect that it is already being distorted by QRPs.

We use descriptive statistics to visualise whether these features are present in PMAR, or at least in the journals analysed. The method is fairly elementary but under the right circumstances rather effective (Gigerenzer, 2004; Gigerenzer & Marewski, 2015) and has been used successfully (Dyckman & Zeff, 2014; Matthes et al., 2015; McCloskey & Ziliak, 1996; Ziliak & McCloskey, 2004b).

The framework looks at the frequency in experimental design, sample type, type of participants, sample selection, sample size, response rate, publication bias, and *P* values. A set of survey questions has also been developed from the theory presented on QRPs in the second section of the paper; for each article, the questions must be answered either with a yes or a no. By combining the survey questions and ratios, the aim is to shed light on whether the publication system of PMAR is susceptible and provides life-support for QRPs. In the following we will explain the relevance of each survey question.

(1) Does the paper replicate a former study to corroborate its findings?

Maniadis et al. (2014) have evidenced that a few independent replications will dramatically increase the chances that original statistical findings are in fact true, as replications decrease the FDR. In this analysis, we follow the broader definition of Dyckman and Zeff (2014) stating that a replication should redo an identical study in the same way but for another but similar sample.

(2) Does the paper disclose information on data availability?

Promoting a data sharing or retention culture would bring transparency and reproducibility to the scientific process, as data sharing or retention would allow other researchers to corroborate or self-correct irreproducible findings (Munafò et al., 2014). It is therefore of interest to investigate whether journals are dedicated to this practice.

(3) Does the paper state the statistical power of the test?

Statistical power is the likelihood that a study will detect an effect when there is an effect to be detected. High statistical power therefore reduces the probability of making a type II error and is directly related to the FDR (Halsey et al., 2015; Ioannidis, 2005; Maniadis et al., 2014).

(4) If the paper mentions power, does it then examine the statistical power of the test?

If the paper does indeed mention statistical power, how many papers do actually examine it? It would be sound scientific conduct if statistical power was investigated *a priori* to determine the required sample size for detecting the expected effect size (Matthes et al., 2015; Simmons et al., 2011).

(5) Does the paper engage in "sign econometrics"?

"Sign econometrics" is about stating the direction of the coefficient but not its size (McCloskey & Ziliak, 1996; Ziliak & McCloskey, 2004b). However, 'sign' is not economically significant unless the magnitude is large enough to matter, and statistical significance does not indicate whether it is large or small (Carver, 1993; Sullivan & Feinn, 2012; Wasserstein & Lazar, 2016; Ziliak, 2016). Low *P* values do not necessarily imply large or more important effects (Wasserstein & Lazar, 2016). Coefficients and effect sizes should therefore be carefully interpreted in relation to the hypothesis under investigation to determine whether the significant statistical findings have scientific or economic relevance. Specifically, an effect size provides quantitative information about the magnitude of the relationship studied; it is therefore much more precise than making a qualitative statement saying that "X increases positively with Y" (Halsey et al., 2015; Hubbard & Lindsay, 2013).

(6) If the paper discusses coefficients, does it then provide confidence intervals for the coefficients?

Confidence intervals also play an important part in interpreting the relevance of coefficients as they convey what a *P* value does not, namely the magnitude and the relative importance of an effect (Nuzzo, 2014). Specifically, by presenting the range within which the true effect size is likely to lie, a confidence interval indicates the uncertainty of a measure. Ziliak and McCloskey (2004a, p. 673) refer to Jeffrey Wooldridge on the importance of these two measures. In his textbook *Introductory Econometrics*, Wooldridge suggests that *"Sign without size, and sign without size without confidence intervals, is mainly beside the point".* By reporting effect sizes and confidence intervals, the statistical interpretation of data emphasises both the importance and precision of the estimated effect size, which, in the end, allows for the importance and relevance of the effect to be judged (Halsey et al., 2015).

(7) Does the paper carefully compare and discuss the findings with previous similar studies?

This question concerns the comparison of sign and effects with previous similar studies, because when findings are confirmed by an independent study, the credibility of the inference made is strengthened. This is also what Maniadis et al. (2014) describe as a solution to the current inferential problem of science; however, it would require that replications are common practice in the scientific field.

3.3 The dataset: published Positivistic Management Accounting Research papers

Our journal selection is focused on journals in which PMAR has been prominently published and it reflects two leading journals according to journal rankings and accounting faculty surveys (Ballas & Theoharakis, 2003; Bonner, Hesford, Van der Stede, & Young, 2006; Lachmann et al., 2017). The period is selected based on the argument that statistical validity has increased over time (Lachmann et al., 2017) and we wish to investigate the prevalence of QRPs in present-day research. We therefore analyse the publication practices of the journals of *Accounting, Organization and Society (AOS)* and *Management Accounting Research (MAR)* in the period 2010-2015 in order to determine if their practices allow QRPs. The approach of this paper follows that of Dyckman and Zeff (2014) in terms of data sampling and selection of articles. We chose to restrict the selection of articles to include only survey and experimental research papers. Consequently, the following types of papers are excluded: (i) theoretical papers, (ii) qualitative papers, (iii) archival papers, (iv) Bayesian estimation methods, and (v) purely explorative papers identifying constructs. This screening leaves us with 38 survey papers and 36 experimental papers, with experimental research being predominantly within AOS (33 out of 36) and survey papers predominantly within MAR (24 out of 38). The papers analysed are presented in appendix A. The sample of articles from two of the leading accounting research journals is argued to be representative of recently published research – an argument also claimed in a similar study by Dyckman and Zeff (2014). The data produced from the articles is a result of direct examination and not from using an electronic search engine. An extract of the coding of the papers is presented in appendix B; by providing the coding of each paper we try to provide transparency on the meta-analysis.

It should be noted that management accounting research is a subset of accounting research, just like financial accounting and auditing research. These accounting research subfields overlap and are not very distinguished, which is why it may be claimed that some of the articles analysed are part of another subset. However, we do not necessarily see this as problematic for the analysis, because the research methods applied will be more or less equivalent due to being published in the same outlets, and the fields draw on the same theoretical base.

4. Analysis

The structure of the analysis follows the structure of the framework presented and discussed in the previous section. It is important to stress that it is neither the intention nor the purpose of this article to claim researcher malfeasance, but to analyse and discuss our current publication system and the related potential danger of QRPs. An extract of the data material for the figures and tables is provided in appendix B. The results for survey studies are presented in Figure 2 and Table 2 while the results for experimental research are presented in Figure 3 and Table 3.

4.1 Results from survey studies

The results indicated a presence of a positive publication bias in survey research, as 24 out of 38 articles confirmed more than 70 percent of their hypotheses and 34 percent confirmed all of their hypotheses (13 out of 38 articles). Furthermore, none of the published studies confirmed less than 40 percent of their hypotheses. In total, 71 percent (164 of 232) of all stated hypotheses were found significant and thus claimed to be true.

Concerning data sharing, not a single survey study provided access to the raw data, which unfortunately is not a surprising finding. Sharing raw data in management accounting research is uncommon as raw data tends to be protected due to confidentiality reasons. However, anonymising raw data might solve the confidentiality issues, but raw data also represents a publication value, illustrated by Bedford (2015) and Bedford and Malmi (2015) using the same survey raw data; both studies are published in MAR. Considering the common saying 'publish or perish', it would, in this light, be rather unwise to share raw data as it provides an opportunity for additional publications and hence career advancement.

Another finding is that replications are a rather rare exemption, as only one study claimed to reproduce or corroborate a previous finding of a similar study. Thus, of the 164 significant hypotheses, it would appear from this meta-analysis that not one single hypothesis has been genuinely replicated or replicates a previous empirical finding. The lack of replications is also illustrated by the fact that not a single study carefully compares coefficients and effect sizes with previous empirical findings. The scarcity of replications within AOS and MAR could be attributed to their high-ranking status, by only publishing what they perceive as novel findings. Similarly, by analysing *The Accounting Review* and *Journal of Accounting Research,* a study by Dyckman and Zeff (2014) found a clear lack of replications in financial accounting, which is a situation that appears to be symptomatic for social science as a whole (Gigerenzer & Marewski, 2015; Goodman et al., 2016; Yong, 2012).

So far, it appears that the publication environment for survey studies in PMAR unintentionally incentivise researchers' engagement in QRPs as it provides the space for *P*-hacking and HARKing to occur. This is due to a positive publication bias, a lack of replications and no raw data transparency. The likelihood of a published false-positive to be falsified or self-corrected is therefore, unfortunately, close to zero.

We now look at the flexibility in sampling and the process of analysing and reporting the data. We found that the typical study, 30 percent of survey studies, has a sample size in the interval of 51-100 respondents, while 66 percent are below 200 respondents. Furthermore, the typical sample type is a non-random sample by approximately 74 percent. Van der Stede et al. (2005) claim, as a rule of thumb, that a survey study should at least attain a sample size between 200-300 respondents to ensure an acceptable level of validity, which is higher than approximately 61 percent of the studies in this analysis. Considering the relatively small sample sizes, it would have

been appropriate that the survey studies had estimated their statistical power. However, only 21 percent mention statistical power, and only 11 percent examine the power of the test. This situation is worsened by the effect sizes within management accounting typically being quite small, and the smaller the effect size, the larger the statistical power required (Borkowski, Welsh, & Zhang, 2001; Sullivan & Feinn, 2012).

In theory, a statistical study should *a priori* calculate the sample size needed by reflecting on the expected effect sizes to ensure that the statistical power is high enough for detecting the expected effect sizes. Henri and Journeault (2010, p. 69) follow this best practice: *"In the current study, the sample size is adequate to test the proposed model (n = 303) as well as the ratio of observations per parameter. Furthermore, based on the guidelines of MacCallum, Browne, and Sugawara (1996), this study has adequate statistical power (i.e. 0.93)".* Power analysis can also be used to *post hoc* analyse if a study correctly rejects a hypothesis. Artz, Homburg, and Rajab (2012, p. 454) illustrate this: "As *a valid theoretical rationale exists for expecting a significant positive interaction effect, an important question is whether our sample size is powerful enough to find an existing effect. Therefore, we analysed whether our sample has adequate statistical power to reject the null hypothesis of no effect... We found that we can detect a true effect size of .050 with about 90% power and a true effect size of .037 with about 80% power... The significant interactions (table 4) have effect sizes greater than .050. Therefore, we conclude that if an effect exists, our setting is powerful enough to find it. We still have to reject H2a".* Unfortunately, these two studies appear to represent a rarity in PMAR, and considering the influence of statistical power on the FDR, it is worrying that statistical power receives so little attention from authors and reviewers.

Concerning bias, the response rate for the analysed survey studies typically ranges from 20 to 40 percent, representing 24 of 38 studies, and it is a well-known fact that a low response rate could induce a *non-sampling error* issue and a *non-response bias*. So, when a sample size is secured to be large enough for the statistical power to be satisfactory, any efforts left should be moved to increase the response rates (Van der Stede et al., 2005). A low response rate represents a potential for increased bias in the PPV equation resulting in an increased ratio of false-positives.

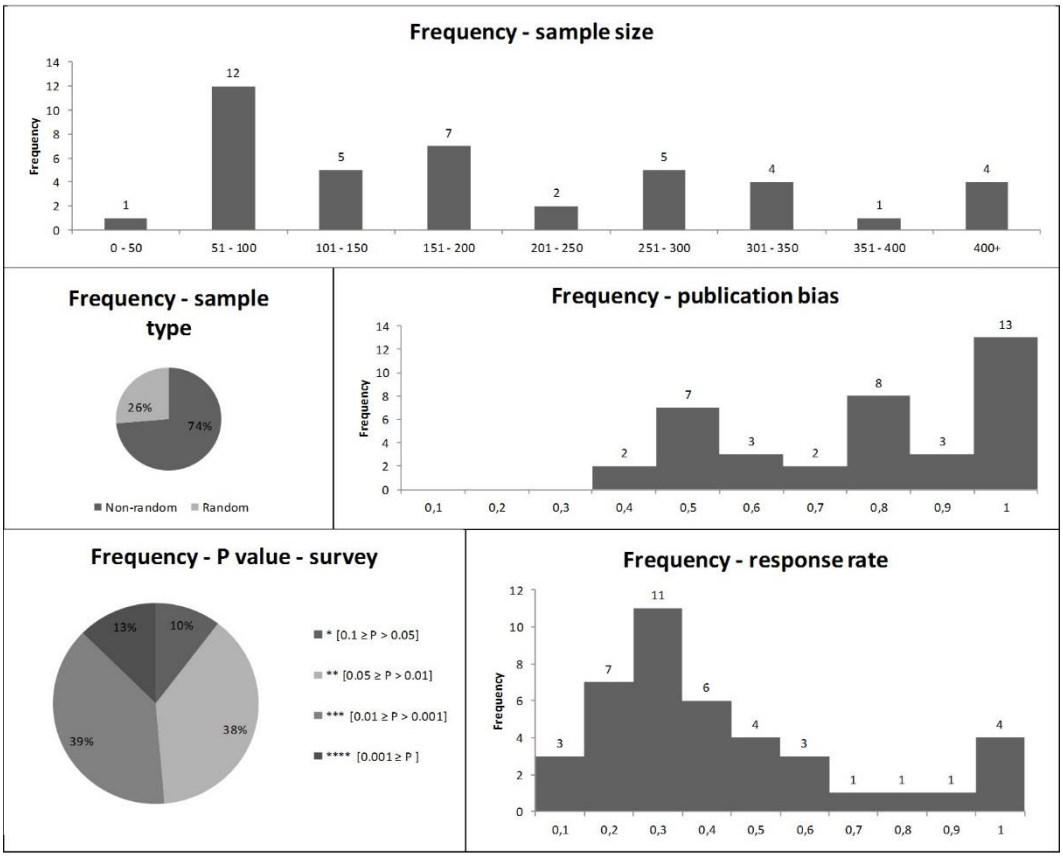


Fig. 2. Dashboard of results from analysing 38 survey studies in AOS and MAR from 2010 to 2015

Table 2. Findings for survey studies in AOS and MAR

Survey Questions	Yes	No
1. Does the paper replicate a former study to corroborate its findings?	3%	97%
2. Does the paper disclose any information on data availability?	0%	100%
3. Does the paper state the statistical power of the test?	21%	79%
4. If the paper mentions power, does it then examine the statistical power of the test?	11%	89%
5. Does the paper engage in "sign econometrics"?	84%	16%
6. If the paper discusses coefficients, does it then provide confidence intervals for the coeffi	5%	95%
7. Does the paper carefully compare and discuss the findings with previous similar studies?	0%	100%

Source: All the full-length papers using tests of statistical significance and published in *AOS or MAR* between 2010 and 2015
N (AOS) = 14, N (MAR) = 24, Total = 38 articles

We now analyse the practice of statistical reporting. We found that 84 percent of the papers engage in "sign econometrics", that is remarking on the direction of the effect but not its size. This is understood as papers basing the relevance of their findings on whether they were significant in the predicted direction instead of discussing the magnitude of the relationship studied. This practice allows *P*-hacking to occur, as *P*-hacking is about searching for significance with no regard to effect sizes. To illustrate the typical ways of reporting statistical evidence in the articles analysed, we present a set of quotes that are representative for the sample.

Bisbe and Malagueño (2012, p. 305) report their results as: *"Panel A in Table 3 displays the results of the causal steps procedure. It shows that SPMS have a positive effect on the strategic decision array variety ($p < 0.01$), which in turn, has a positive effect on organisational performance measured through ROS ($p < 0.05$). Analogously, results show that SPMS have a positive effect*

on the strategic decision array size ($p < 0.01$) as well as on ROA ($p < 0.05$). Overall, these results suggest that, as predicted by H1a, SPMS are positively associated with the comprehensiveness of the strategic decision arrays". This quote illustrates a typical practice of reporting coefficients and effect sizes in tables but refrains from commenting or reflecting on the effect sizes. Another example of "sign econometrics" is from Ho, Wu, and Wu (2014, p. 48): *"Specifically, employees' tenure (EMP_TENURE) is positively and significantly associated with customer satisfaction (0.02, $t = 2.97$, $p < 0.01$ in Model 1 and 0.02, $t = 3.19$, $p < 0.01$ in Model 2), which suggest that senior salespeople provide more satisfying service and have earned greater trust and loyalty from their customers".* The researchers refrain from discussing the practical or scientific relevance of their findings despite their coefficients appearing to be very small and perhaps rendering the results irrelevant.

However, not all papers omit discussions of effect sizes. An example is Guerreiro, Rodrigues, and Craig (2012, pp. 493-494) who apply an *odds ratio* approach to differentiate between significant predictors by claiming that the most important predictors are those that change the odds of the outcome the most: *"As column Exp(B) of Table 4 reveals, firms with listed parent companies are 12.93 times more likely to adopt IFRS voluntarily than are firms with unlisted parent companies".* This study found the calculated odds ratios to range from 0.173 to 12.925 with all predictors being statistically significant thereby allowing researchers to judge the relative relevance between predictors.

The survey studies analysed argued their relevance on significance and there was an almost complete lack of discussing coefficients and effect sizes. A practice of "sign econometrics" is unfortunate, not only by allowing *P*-hacking, but also because small *P* values do not imply importance or relevance. Instead careful consideration of coefficients and effect sizes would be not only a prerequisite for claiming scientific or economic significance but also discourage *P*-hacking (Wasserstein & Lazar, 2016; Ziliak, 2016; Ziliak & McCloskey, 2004b). The focus on significance is further evidenced by a complete disregard for confidence intervals despite confidence intervals being superior to *P* values in every way.

Concluding on the findings, the publication practice related to survey studies give rise to concern. The findings indicate that the PMAR publication system on survey studies is susceptible to QRPs and unintentionally incentivise researchers to engage in QRPs. No evidence is found that the publication practice of PMAR is discouraging the activities of *P-hacking* or HARKing. In terms of publication practices and statistical reporting, PMAR therefore appears to be following in the footsteps of social science by allowing the hypothetico-deductive method to be distorted by QRPs. However, methodological critique is not unknown to survey studies (Van der Stede et al., 2005), and many see experiments as a response to this critique; experiments are therefore becoming increasingly popular.

The next subsection investigates if the publication practice of experimental research in PMAR is more resilient to QRPs, which unfortunately has not been the case for other scientific fields (Camerer et al., 2016; Matthes et al., 2015; Simmons et al., 2011).

4.2 Results from experimental studies

Our results indicate an even more consistent positive publication bias than observed within survey studies, as 24 out of 36 articles confirm 100 percent of their hypotheses. In total, 86 percent (99 of 115) of all stated hypotheses were found to be statistically significant and claimed to be true. Concerning data availability, it is the same case as for survey studies; not a single study provided access to raw data, although it is quite common for studies to provide a detailed experimental design guide either as an appendix or as online supplementary material. A guide gives researchers the opportunity to replicate experimental findings, however only 6 percent (2 out of 38) claimed to replicate former empirical findings in a new setting by a new experimental design, thereby not representing a 'genuine' replication. This would require the precise same experimental design to be used (Dyckman & Zeff, 2014; Moonesinghe et al., 2007). As noted before, the reason for the lack of replications could, again, be that the prestigious journals prefer novel findings and therefore are reluctant to publish replications. However, another factor is possibly due to experiments within management accounting being in the making thus limiting the number of studies available for replication. Nonetheless, a lack of replications is unfortunate for the accumulation of knowledge, as evidenced by Maniadis et al. (2014) who demonstrate that the FDR diminishes drastically even after a few replications. The lack of replications therefore represents a concern for PMAR. For example, economic research has found itself limited in its ability to reproduce previous experimental findings, as Camerer et al. (2016) tried to replicate 18 experimental studies published in the *American Economic Review* and in the *Quarterly Journal of Economics*; they only found a significant effect in the same direction as in the original study for 11 replications, and on average the replicated effect size was only 66 percent of the original.

Based on these findings, we draw the same conclusion for the publication environment for experimental studies as we did for survey studies, namely, that it provides the potential to engage in *P*-hacking and HARKing without researchers having to fear being 'caught'. This is because of a high publication bias, a lack of replications and no transparency on raw data resulting in the likelihood of false-positive to be falsified or self-corrected being, once again, close to zero.

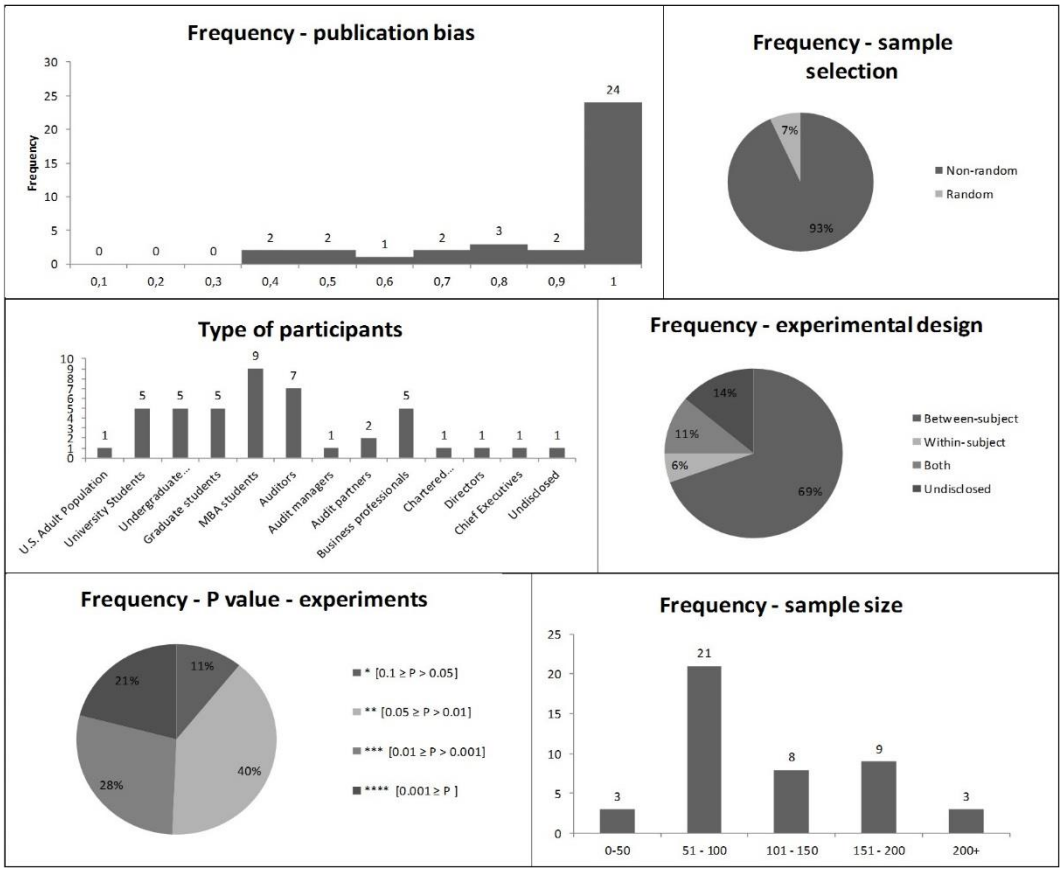


Fig. 3. Dashboard of results from analysing 36 experimental studies in AOS and MAR from 2010 to 2015

Table 3. Findings for experimental studies in AOS and MAR

Survey Questions	Yes	No
1. Does the paper replicate a former study to corroborate its findings?	6%	94%
2. Does the paper disclose any information on data availability?	0%	100%
3. Does the paper state the statistical power of the test?	8%	92%
4. If the paper mentions power, does it then examine the statistical power of the test?	0%	100%
5. Does the paper engage in "sign econometrics"?	94%	6%
6. If the paper discusses coefficients, does it then provide confidence intervals for the coeffi	0%	100%
7. Does the paper carefully compare and discuss the findings with previous similar studies?	0%	100%

Source: All the full-length papers using tests of statistical significance and published in AOS or MAR between 2010 and 2015
N (AOS) = 33, N (MAR) = 3, Total = 36 articles

We now look at the flexibility in sampling and the process of analysing and reporting the data. We found that 93 percent of the studies use a non-random sample. The typical participant is a student, as university students take part in more than 40 percent of the experiments (15 of 36), while MBA students account for 25 percent (9 of 36). It might be called into question whether these types of participants represent the population of interest. On the positive side, there is also a noteworthy high frequency of auditors (9 of 36).

Concerning the sample sizes, it is surprising that 66 percent of the studies analysed have a sample size below 100 participants; an *a priori* statistical power test would have been appropriate to confer that the study was large enough to detect the expected effect size. However, only eight percent of the studies mention statistical power and not a single study conducts an *a priori* analysis of the sample size needed, and also, no study conducts a *post hoc* statistical power test to uncover

if the study failed to reject a H_0 hypothesis. Perhaps this is unsurprising, considering that 86 percent of all hypotheses were found to be significant. The ratio of 86 percent could be claimed to represent an impressive foresight in theoretical forecast, but it is more likely to represent some degree of *P*-hacking or HARKing if the publication environment permits it (Leung, 2011). It is, however, unfortunate that the authors and reviewers put so little emphasis on the calculation of statistical power tests.

We now analyse the practice of statistical reporting: 94 percent of the experimental studies were found to engage predominantly in "sign econometrics", demonstrated by a low emphasis on coefficients and effect sizes despite a wide range of effect size measures for experimental research being available, e.g. Cohen's d, Odds Ratio, Relative Risk or Risk Ratio, Pearson's r correlation, r^2 coefficient of determination and Glass' Δ (Sullivan & Feinn, 2012).

To illustrate the typical way of reporting statistical evidence, we, again, draw on a selection of quotes. However, the reporting of statistical evidence is very similar to that in survey research and we therefore only two representative quotes. For example, Perreault and Kida (2011, p. 542) report the statistical findings as: *"Compared to an auditor using a cooperative communication style, those that negotiated with the contentious auditor liked the auditor less (3.13 vs. 4.72; t = 6.58; p < .001), were less happy with the auditor (3.36 vs. 4.84; t = 5.68; p < .001), were more frustrated with the auditor (4.44 vs. 3.59; t = 3.21, p < .001) and were more angry with the auditor (3.46 vs 2.67; t = 3.20; p < .001). As a result hypothesis four is supported".* While Chen and Tan (2013, p. 221) only focused on the *P* value: *"We find a significant exposure effect on the change in participants' third quarter earnings estimates in the absence of the performance cue (p = 0.04), and an insignificant exposure effect in its presence (p = 0.29) ... In Sum Hypotheses 1 and 2 are supported for both credibility and earnings estimates measures".*

Once again, we observe that experimental studies tend to focus on the *P* values and related significance while refraining from discussing coefficients or effect sizes. In addition, not a single paper calculated and presented confidence intervals as a part of the discussion. Reviewers' reluctance to require a discussion of coefficients is not only unfortunate but also 'poor' statistical reporting, as the *American Statistical Association*, on the matter of QRPs, argues that scientific importance is not solely a matter of significance, but just as much of confidence intervals, coefficients and effect sizes (Wasserstein & Lazar, 2016).

As for survey studies, we also found reason for concern as regards experimental research. On various parameters, experimental research is even more indicative of a publication practice that favours QRPs, as we found a stronger publication bias, smaller sample sizes, bias in use of participants, and an even stronger tendency to engage in "sign econometrics". Based on these observations, we argue that the publication practices of experimental research in PMAR unintentionally encourage the phenomenon of QRPs which is why we are concerned about the size of the ratio of false-positives. As such, the theoretical advantage of experimental research in

making causal claims appears to be offset by the likelihood of QRPs taking place. On the bright side, the method allows easier replication and hence the detection of false-positives as it would only require our outlets to begin to publish replications.

4.3 Discussion

The meta-analysis evidences that it is reasonable to argue that the publication practices of PMAR is susceptible to QRPs and, as such, it is likely that our field is also being distorted by this phenomenon. This is based on the identification of a set of current characteristics in the PMAR publication system where we found a positive publication bias, a non-random sampling, a lack of replications, no data sharing, no statistical power tests, and a common practice of "sign econometrics". From the meta-analysis, we highlight the following key ratios:

- High positive publication bias for survey studies [164 of 232 hypotheses confirmed, 71%] and experimental studies [99 of 115 hypotheses confirmed, 86%].
- Non-random sampling for surveys [74%] for experimental [93%].
- Sample size for survey studies [66% below 200 respondents and response rate of 20-40%] for experimental studies [typically 51-100 participants, and the typical participant is a student].
- A complete lack of replications, no data sharing and a negligence of statistical power tests [either *a priori* or *post hoc].*
- Common engagement in "sign econometrics" for survey studies [84%] and experimental studies [94%], while disregarding presentation and discussion of confidence intervals for survey studies [5%] and experimental studies [0%].
- A complete lack of comparing findings [coefficients and/or effect sizes] with previous, similar studies.

Research has evidenced the existence of QRPs in the social sciences and the natural sciences, and we found no evidence suggesting that QRPs have not taken foothold in PMAR. It is therefore rational to expect that the false-positive ratio is significantly higher than the conventional ratio of five percent; however, the precise extent of false-positives is impossible to verify without replications. We therefore expect PMAR to be producing research findings when they should not be produced, supporting a major concern, raised by Ioannidis (2005), for statistical research as a whole.

Our findings outline a different reality for the validity of PMAR than the one painted by Lachmann et al. (2017). However, we very much agree that in terms of those characteristics investigated by Lachmann et al. (2017) the validity of PMAR has been increasing over the last four decades. On the other hand, we argue that the validity of PMAR cannot be discussed without taking QRPs into account. It is therefore not enough to discuss only the criteria of internal validity, external validity, construct validity and statistical validity[13]. This will not provide a holistic picture

[13] Internal validity is coded in terms of *time frame*. External validity is coded in terms of *type of sample* and *primary occupation of participants*. Construct validity is coded in terms of *number of measures for construct validation, number of reliability measures, type of dependent variables, and number of data*

of the validity of PMAR, i.e. the ability to replicate original findings. If we are right about the PMAR publication practices of, then it amounts to a serious problem for the reliability of PMAR in terms of whether research findings can be considered as 'true' (Ioannidis, 2005; Ioannidis & Doucouliagos, 2013; Young, Ioannidis, & Al-Ubaydli, 2008); also, it indicates a paradox: To live up to the positivistic image of 'pure science' published in academic journals, researchers find themselves – ironically – transgressing this very ideal (Butler et al., 2017).

Genuine replications would have provided hard evidence of the existence of QRPs; unfortunately, this is a type of study that is underappreciated in most social sciences, including PMAR. Self-correction is a scientific feature that seems long forgotten in social sciences as a whole.

To put it bluntly, the publication practice of PMAR can be understood from a column written by Andrew Gelman and Erik Loken where they compare the practice of publishing in science with the subprime mortgage crisis in the United States. Their column is framed 'The AAA Tranche of Subprime Science' and was published in Change:

> *"The first step is statistical significance. Out of the primordial soup of all possible data analyses, the statistically significant comparisons float to the top. They represent the high-certainty statements selected out of the many less-reliable claims. The second step is publication in a scientific journal, ideally a high prestige outlet... -But, if not a top journal any outlet will do. The convention is to treat published claims as true unless demonstrated otherwise. The two-step process – first the achievement of statistical significance, then publication – corresponds with the movement of scientific hypothesis from the hazy zone of uncertain speculation to presumed certainty".*
>
> (Gelman & Loken, 2014a, p. 51).

They describe a system where the role of statistical significance and the peer review process represent a seal of approval for scientific claim (Bedeian, 2004; Dyckman & Zeff, 2014; Ohlson, 2015). However, this system is allegedly challenged as neither statistical significance nor the peer review process, as currently practiced, work quite as intended (Bamber et al., 2000; Banks, Rogelberg, et al., 2016; Garud, 2015; Maniadis et al., 2014; Starbuck, 2016).

The meta-analysis indicates that the publication practice of PMAR contains similarities with scientific fields that we know are being distorted by QRPs, as data analysis or scientific inference seem to have been reduced to a mechanical 'bright-line' rule, i.e. $P < 0.05$, when justifying for scientific claim. For example, only a very few studies argue for the relevance of their empirical findings on other information than the decisive factor of the *P* value being below 0.05, while not a single study counter-argues the relevance of a significant hypothesis due to a too small effect size rendering the finding irrelevant. Such a practice has the potential of leading to erroneous beliefs and poor decision-making (Wasserstein & Lazar, 2016). A use of 'significance',

sources. Statistical validity is coded in terms of particular tests of *multicollinearity, omitted variable bias, simultaneity bias, self-selection bias, heteroscedasticity, outliers* and so on (Lachmann et al., 2017).

as *the* license for claiming a scientific finding has the potential of leading to considerable distortion in the scientific process of the hypothetico-deductive method. An empirical finding does not immediately become 'true' on one side of the divide and 'false' on the other (Wasserstein & Lazar, 2016) and to quote Sir R. A. Fisher:

> *"No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."*
>
> (Fisher, 1956, p. 42).

Relying on significance provides the perfect incentives for *P*-hacking and HARKing, as invariably there is a data analysis path which leads to significance even in the absence of an underlying effect, and – if not – a researcher can always change hypothesis (Loken & Gelman, 2017). The publication practices of PMAR therefore are in danger of upending the assumption about the number of studies required to produce a false-positive finding, thereby questioning the validity of our knowledge base.

The *American Statistical Association* elaborates on the meaning of *P* values arguing that a *P* value does not measure the probability that a studied hypothesis is true or even relevant (Wasserstein & Lazar, 2016). Mainly reporting and concluding on significance and directional effects instead of arguing for scientific significance or economic significance therefore in reality and in essence constitute statistical misbehaviour (Dyckman & Zeff, 2014; Evans, Feng, Hoffman, Moser, & Van der Stede, 2015; Gigerenzer, 2004; Gigerenzer & Marewski, 2015; Lindsay, 1994; Ziliak & McCloskey, 2004b).

The purpose and ambition of PMAR are to develop reliable and valid causal explanations of management accounting phenomena and thus to draw inferences from a sample of specific observations to the general (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2014). The prestigious academic journals of *Accounting, Organizations and Society* and *Management Accounting Research* should be beacons of scientific integrity and reliability. It is therefore problematic for PMAR as a whole when we find the publication practices of these two journals to be susceptible and perhaps even (unintentionally) incentivising QRPs.

In the next section, we will put forward suggestions for the adaptation of the publication practices of PMAR towards becoming more resilient towards QRPs hence strengthening the credibility of our scientific field.

5. Conclusion and suggestions

Our findings indicated that the current research traditions of PMAR rendered it possible for QRPs to take root. Based on our findings, we question the reproducibility of PMAR and therefore expect the false-positive ratio to be well above the conventional five-percent ratio. As a result, we would expect PMAR to produce research findings when they should not, which was

the main concern of QRPs and its distortion on the hypothetico-deductive method first raised by Ioannidis (2005).

5.1 The establishment of a bad equilibrium in PMAR?

Any scientific field should strive towards becoming an accumulative, iterative, self-correcting endeavour, where mistakes, such as false-positives, are a normal short-term effect of a long-term process of accumulating knowledge. However, for this to be the case, a publication system that does not incentivise researchers to stray from this path and indulge in activities of QRPs is required. A scientific field should avoid "sign econometrics", positive publication bias, lack of data sharing and a lack of replications. However, it appears that the 'pressure to publish' combined with the competitive advantage of QRPs have created a situation where:

> *"... it is no longer worth questioning questionable research practices (QRPs). If these practices are widespread, then we have all become prisoners of a system that we have created but are unable to free ourselves from. Macro expectations and incentives have transformed into micro motives... Problematic practices that violate the percepts of basic methods appear now to be accepted as normal"*
>
> (Garud, 2015, p. 452).

We believe that the dysfunctionality of the current publication system in relation to QRPs has permitted a bad equilibrium to unfold (Butler et al., 2017).

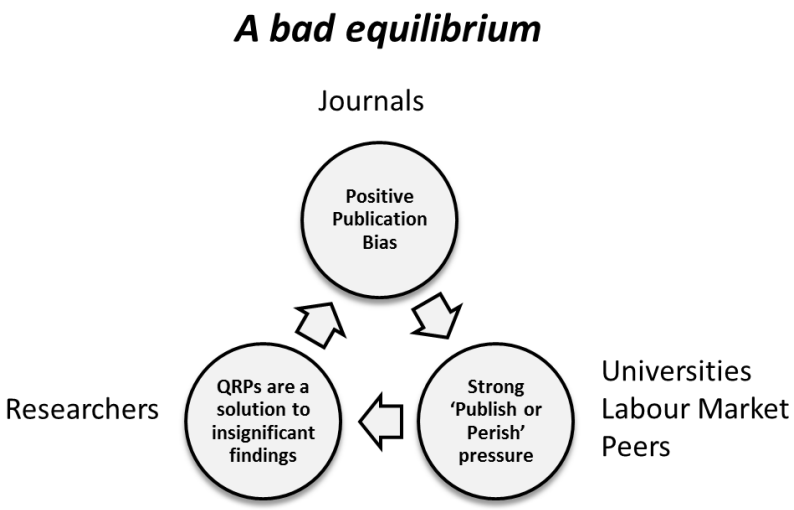


Fig. 4. A self-sustaining equilibrium

The equilibrium is sustained due to the 'publish or perish' pressure and the fact that QRPs are a viable solution to insignificant findings. In other words, in a scientific field where *P* values prevail, researchers will tend to chase the asterisk that signal 'statistical significance' as it is the catalyst for justifying theoretical assertions or scientific relevance (Gelman & Loken, 2014a). However, a publication system where "statistical significance" + "publication" = "truth", provides the potential for *P*-hacking or HARKing to be the solution to insignificant findings. The current publication system of PMAR appears to be a system that unintentionally 'promotes' instead of

'intentionally' hinders false-positives (Ohlson, 2015). If this situation does not change, in the long run it will unquestionably lead to devastating consequences for society, as it may lead researchers, policymakers and funding agencies down false paths, stifling and potentially eroding scientific progress while wasting society's resources (Simonsohn et al., 2014).

Nature and *Science* have united in an effort to break this bad equilibrium by bolstering statistical research. *Science* is adding an extra round of statistical checks to its peer-review process, which is conducted in collaboration with the *American Statistical Association* and a new *Statistics Board of Reviewing Editors* (SBoRE) consisting of seven expert statisticians (McNutt, 2014). *Nature,* on the other hand, has developed a *statistical checklist*[14] to improve the statistical robustness of results and, furthermore, employs statistical consultants to the review process of certain papers; this is done at the discretion of the editors or, if suggested, by the referees (Van Noorden, 2014).

We hope that in the future PMAR will strive for the same robustness against QRPs. If incorrect causal claims are generalised to practice, it will not only erode scientific progress but be devastating to the society that we as researchers claim to serve.

In the next section, we will present three suggestions for breaking the bad equilibrium.

5.2 How to break the cycle?

The trigger of the bad equilibrium is arguably the "journals" (Butler et al., 2017), so for any solution to be viable, this is where the problem must be solved. At first glance, increasing transparency in mitigating for QRPs would be the natural thing to do as the issue is either misreporting or lack of reporting. In addition, a system built to appreciate reports of magnitudes instead of "sign econometrics" would include resistance towards *P*-hacking and HARKing; however, this disincentive might not be strong enough. We therefore propose three solutions for implementation by journals; and we will evaluate how realistic each of these solutions is and if their impact is forceful enough to counter QRPs.

Solution 1

Researchers must submit research projects instead of final papers. Research projects contain research question(s), theory and hypotheses development along with a detailed research design. Data collection and results should be produced after the project has been accepted for publication ensuring that also non-significant results will be published. The approach of this solution would completely discourage any engagement in QRPs; this approach resembles the 2017 *Journal of Accounting Research (JAR)* experiment on "Registered Reports". Whether the approach is realistic is questionable, it would change the current scientific system drastically, but

[14] http://www.nature.com/authors/policies/checklist.pdf

it is bound to break the cycle of the bad equilibrium. It will be interesting to follow the JAR experiment.

Solution 2

The second proposal concerns the publishing of replications. The knowledge that researchers are likely to reproduce from previous findings would, in the long run, trigger concerns for reputation and hence discourage QRPs. In addition, a few independent replications would also dramatically increase the probability that the original finding is true (Maniadis et al., 2014). On the other hand, we would risk overemphasizing replications thereby shifting attention away from exploiting innovations and novel findings and hence slowing down the progress of research. Furthermore, who would undertake the responsibility of doing replications if it indicates lack of creativity and originality? It might be unrealistic to expect major journals to change their strategy and actively publish replications. However, it might be an opportunity for "second-tier" journals to "move up" if the broader academic public would appreciate replications.

Solution 3

The third proposal is a system of "variance investigation", which is defined as a system where a paper might be lifted from the archives for replication purposes. If researchers are aware that this might happen and if it is not too unlikely, it would trigger reputation concerns for the researcher if the replication fails to corroborate the original findings. This would introduce a level of inertia for a researcher to engage in QRPs. Also, this system is realistic as journals might use it as a seal of 'quality' thus enabling them to benefit from it.

None of these proposals should stand alone, and we therefore would like to point out the following for consideration independently of the above proposals. First, journals must make sure that their reviewers require authors to use more facets of the statistical toolbox in the argumentation of their results than solely judging on statistical significance. In this way, they would also encourage a discussion of magnitudes when claiming for scientific relevance (Dyckman & Zeff, 2014). Second, disclosing the number of hypotheses explored, all data collection decisions, all statistical analyses conducted and all *P* values computed is a precondition (Wasserstein & Lazar, 2016). These papers should be branded as *'P certified, not P hacked'* including the following wording: "*We report how we determined our sample size, all data exclusions (if any), all manipulations and all measures in the study*" (Nuzzo, 2014). Third, in preventing publication bias compromising our knowledge base, it is pivotal that research outlets move away from the seemingly positive publication bias: by appreciating the inherent knowledge in insignificant findings and by not encouraging their exclusion in the review process (Banks, O'Boyle, et al., 2016). Fourth, journals should develop a data sharing culture as it would give researchers the opportunity to verify original analyses. Commitment to complete transparency as regards original data is an important property of 'good' science, and journals should therefore develop policies

on how to handle data sharing (Munafò et al., 2014; Nosek et al., 2015; Tenopir et al., 2011). Mitigating confidentiality issues trough anonymising raw data ought to be possible; this is already done in medical research on confidentiality concerns of private data. Fifth, a theme also present in previously proposed solutions is replications, being a fundamental requirement for the creation of rigorous theories and opposing the accumulation and dissemination of false knowledge (Merton, 1942, 1973). By pushing for replications, we allow science to be self-correcting and we would conform to a core requirement for claiming causality argued by David Hume in '*Enquiries of Human Understanding*' from 1748:

> *"Even after one instance or experiment, where we have observed a particular event to follow upon another, we are not entitled to form a general rule, or foretell what will happen in like cases; it being justly esteemed an unpardonable temerity to judge of the whole course of nature from one single experiment, however accurate or certain."*
>
> (Hume, 1975, p. 74)

By following these guidelines and considering the three solutions proposed, it should help ensure that a 'bright-line' rule does not steer the publication process and, in the end, ensure that future policy or business decisions in society are not indirectly based on whether a P value passed a certain threshold. Just because we do not necessarily enjoy the methodological precision as natural sciences does (Aguinis & Edwards, 2014), it does not mean that we cannot strive for the same rigour in statistical method, which is a basic prerequisite for theoretical progress and the accumulation of knowledge.

Acknowledgements

I would like to thank the participants at the 10th Conference on New Directions in Management Accounting, Brussel 2016, and in particular Frank Moers, Teemu Laine, Tuomas Korhonen, Rafael Heinzelmann, Morten Jakobsen and Hanne Nørreklit for their comments and suggestions.

Appendix A

List of the seventy-four articles in the sample

Management Accounting Research (2010 – 2015) – Survey research

- Abernethy, Bouwens, and Lent – Leadership and control system design (2010)
- King, Clarkson, and Wallace – Budgeting practices and performance in small healthcare businesses (2010)
- Hall – Do comprehensive performance measurement systems help or hinder managers' mental model development? (2011)
- Lee and Yang – Organization structure, competition and performance measurement systems and their joint effects on performance (2011)
- Weißenberger and Angelkort – Integration of financial and management accounting systems: The meditating influence of a consistent financial language on controllership effectiveness (2011)
- Burkert, Fischer, and Schäffer – Application of the controllability principle and managerial performance: The role of perceptions (2011)
- Windolph and Moeller – Open-book accounting: Reasons for failure of inter-firm cooperation (2012)
- Hartmann and Slapničar – The perceived fairness of performance evaluation: The role of uncertainty (2012)
- Speckbacher and Wentges – The impact of family control on the use of performance measures in strategic target setting and incentive compensation: A research note (2012)
- Caglio and Ditillo – Opening the black box of management accounting information exchanges in buyer-supplier relationships (2012)
- Bisbe and Malagueño – Using strategic performance measurement systems for strategy formulation: Does it work in dynamic environments?
- Burkert and Lueg – Differences in the sophistication of Value-based Management – The role of top executives (2013)
- Dekker, Sakaguchi, and Kawai – Beyond the contract: Managing risk in supply chain relations (2013)
- Ding, Dekker and Groot – Risk, partner selection and contractual control in interfirm relationships (2013)
- Pondeville, Swaen, and De Rongé – Environmental management control systems: The role of contextual and strategic factors (2013)
- Marginson, McAulay, Roush, and Zijl – Examining a *positive* psychological role for performance measures (2014)
- Ylinen and Gullkvist – The effects of organic and mechanistic control in exploratory and exploitative innovations (2014)
- Speklé and Verbeeten – The use of performance measurement systems in the public sector: Effects on performance (2014)
- Ming Chong and Mahama – The impact of interactive and diagnostic uses of budgets on team effectiveness (2014)
- Janke, Machlendorf and Weber – An exploratory study of the reciprocal relationship between interactive use of management control systems and perception of negative external crisis effects (2014)

- Su, Baird and Schoch – The moderating effect of organisational life cycle stages on the association between the interactive and diagnostic approaches to using controls with organisational performance (2015)
- Bedford – Management control systems across different modes of innovation: Implications for firm performance (2015)
- De Baerdemaeker and Bruggeman – The impact of participation in strategic planning on managers' creation of budgetary slack: The mediating role of autonomous motivation and affective organisational commitment (2015)
- Lisi – Translating environmental motivations into performance: The role of environmental performance measurement systems (2015)

Accounting, Organization and Society (2010 – 2015) – Survey research

- Henri and Journeault – Eco-control: The influence of management control systems on environmental and economic performance (2010)
- Veen-Dirks – Different uses of performance measures: The evaluation versus reward of production managers (2010)
- Grafton, Lillis, and Widener – The role of performance measurement and evaluation in building organizational capabilities and performance (2010)
- Bol and Moers – The dynamics of incentive contracting: The role of learning in the diffusion process (2010)
- Herda and Lavelle – The effects of organizational fairness and commitment on the extent of benefits big four alumni provide their former firm (2011)
- O'Connor, Vera-Muñoz, and Chan – Competitive forces and the importance of management control systems in emerging-economy firms: The moderating effect of international market orientation (2011)
- Fayard, Lee, Leitch, and Kettinger – Effect of internal cost management, information systems integration, and absorptive capacity on inter-organizational cost management in supply chains (2012)
- Artz, Homburg, Rajab – Performance-measurement system design and functional strategic decision influence: The role of performance-measures properties (2012)
- Guerreiro, Rodrigues, and Craig – Voluntary adoption of International Financial Reporting Standards by large unlisted companies in Portugal – Institutional logics and strategic responses (2012)
- Fullerton, Kennedy, Widener – Management accounting and control practices in a lean manufacturing environment (2013)
- Ho, Wu, and Wu – Performance measures, consensus on strategy implementation, and performance: Evidence from the operational-level of organizations (2014)
- Mahlendorf, Kleinschmit, and Perego – Relational effects of relative performance information: The role of professional identity (2014)
- Arnold and Artz – Target difficulty, target flexibility, and firm performance: Evidence from business units' targets (2015)
- King and Clarkson – Management control system design, ownership, and performance in professional service organisations (2015)

Management Accounting Research (2010 – 2015) – Experimental research

- Kelvin Liu and Leitch – Performance effects of setting targets and pay-performance relations before or after operations (2013)
- Denker, Schwartz, Ward, and Young – Voluntary disclosure in a bargaining setting: A research note (2014)
- Cheng and Coyte – The effects of incentive subjectivity and strategy communication on knowledge-sharing and extra-role behaviours (2014)

Accounting, Organization and Society (2010 – 2015) – Experimental research

- Koch and Schmidt – Disclosing conflicts of interest – Do experience and reputation matter? (2010)
- Schultz Jr., Bierstaker, and O'Donnell – Integrating business risk into auditor judgment about the risk of material misstatement: The influence of a strategic-system-audit approach (2010)
- Knechel, Salterio, Kochetova-Kozloski – The effect of benchmarked performance measures and strategic analysis on auditors' risk assessments and mental models (2010)
- Cianci and Kaplan – The effect of CEO reputation and explanations for poor performance on investors' judgement about the company's future performance and management (2010)
- O'Donnel and Prather-Kinsey – Nationality and differences in auditor risk assessment: A research note with experimental evidence (2010)
- Norman, Rose, and Rose – Internal audit reporting lines, fraud risk decomposition, and assessments of fraud risk (2010)
- Gibbins, McCracken, and Salterio – The auditor's strategy selection for negotiation with management: Flexibility of initial accounting position and nature of the relationship (2010)
- Cardinaels and Veen-Dirks – Financial versus non-financial information: The impact of information organization and presentation in a Balanced Scorecard (2010)
- Seifert, Sweeney, Joireman, and Thornton – The influence of organizational justice on accountant whistleblowing (2010)
- Jackson, Rodgers, Tuttle – The effect of depreciation method choice on asset selling prices (2010)
- Ranking and Sayre – Responses to risk in tournaments (2011)
- Norman, Rose, and Suh – The effects of disclosure type and audit committee expertise on Chief Audit Executives' tolerance for financial misstatements (2011)
- Gaynor, McDaniel, and Yohn – Fair value accounting for liabilities: The role of disclosures in unravelling the counterintuitive income statement effect from credit risk changes (2011)
- Tan and Koonce – Investors' reactions to retractions and corrections of management earnings forecasts (2011)
- Brüggen and Luft – Capital rationing, competition, and misrepresentation in budget forecasts (2011)
- Perreault and Kida – The relative effectiveness of persuasion tactics in auditor-client negotiations (2011)
- Chen, Kelly, Salterio – Do changes in audit actions and attitudes consistent with increased auditor scepticism deter aggressive earnings management? An experimental investigation (2012)
- Church, Hannan, and Kuang – Shared interest and honesty in budget reporting (2012)
- DeZoort, Holt, and Taylor – A test of the audit reliability framework using lenders' judgements (2012)

- Chang, Chen, and Trotman – The effect of outcome and process accountability on customer-supplier negotiations (2013)
- Chen and Tan – Judgement effects of familiarity with an analyst's name (2013)
- Rose, Mazza, Norman, and Rose – The influence of director stock ownership and board discussion transparency on financial reporting quality (2013)
- Messier Jr., Quick, and Vandervelde – The influence of process accountability and accounting standard type on auditor usage of a status quo heuristic (2014)
- Newman – An investigation of how the informal communication of firm preferences influences managerial honesty (2014)
- Brown, Fisher, Sooy, and Sprinkle – The effect of rankings on honesty in budget reporting (2014)
- Newman and Tafkov – Relative performance information in tournaments with different prize structures (2014)
- Fanning and Piercey – Internal auditors' use of interpersonal likability, arguments, and accounting information in a corporate governance setting (2014)
- Managing audits to manage earnings: The impact of diversions on an auditor's detection of earnings management (2015)
- Lachmann, Stefani, and Wöhrmann – Fair value accounting for liabilities: Presentation format of credit risk changes and individual information processing (2015)
- Gopalakrishnan, Libby, Samuels, and Swenson – The effect of cost goal specificity and new product development process on cost reduction performance (2015)
- Arnold and Gillenkirch – Using negotiated budgets for planning and performance evaluation: An experimental study (2015)
- Agoglia, Hatfield, and Lambert – Audit team reporting: An agency theory perspective (2015)
- Church, Peytcheva, Yu, and Singtokul – Perspective talking in auditor-manager interactions: An experimental investigation of auditor behaviour (2015)

Appendix B

Extract of data for the empirical analysis

#	Journal	Authors	Volume (Year)	Study type	Replication	Information disclosed for replication?	Sample type	N	Sample size	Response rate	# hypothesis	Hypothesis	Result (0/1)	P Value	Disclose data availability	Mentioning statistical power	Examining the statistical power of the test	Publication bias	Engage in "sign econometrics"	Provide Confidence Intervals	Carefully compare and discuss findings with previous similar studies?
1	MAR	De Baerdemae	29 (2015)	Cross-sectional	No	No	Random	2045	249	12%	5	1	0		No	No	No	0.8	Yes	No	No
												2	1	***							
												3	1	***							
												4	1	***							
												5	1	**							
2	MAR	Eleonora Lisi	29 (2015)	Cross-sectional	No	No	Non-random	443	91	21%	5	1a	1	***	No	Yes	No	1	Yes	No	No
												1b	1	**							
												1c	1	***							
												2a	1	***							
												2b	1	*							
3	MAR	Bedford	28 (2015)	Cross-sectional	No	Yes	Random	911	421	46%	10	1	1	**	No	No	No	0.5	Yes	No	No
												2	1	**							
												3	1	**							
												4	0								
												5	0								
												6	0								
												7	1	**							
												8	0								
												9	1	**							
												10	0								
4	MAR	Su, Baird and S	26 (2015)	Cross-sectional	No	Yes	Random	1000	343	34%	4	1	1	**	No	No	No	1	Yes	No	No
												2	1	**							
												3	1	**							
														*							
												4	1	**							
														**							
5	MAR	Janke, Mahler	25 (2014)	o -point Longitudi	No	(Yes)	Random	722	361	50%	2	1	1	***	No	No	No	1	Yes	No	No
								850	277	33%		2	1	**							
6	MAR	Ming Chong a	25 (2014)	Cross-sectional	Yes	(Yes)	Non-random	2000	186	9%	5	1a	1	***	No	No	No	0.6	Yes	No	No
												1b	0								
												2a	1	**							
												2b	0								
												3	1	****							
7	MAR	Speklé and Ve	25 (2014)	Cross-sectional	No	No	Non-random	97	97	100%	3	1	1	**	No	No	No	0.67	Yes	No	No
												2	1	*							
												3	0								
8	MAR	Marginson, M	25 (2014)	Cross-sectional	No	No	Non-random	284	98	35%	5	1	1	***	No	No	No	1	Yes	No	No
												2	1	**							
														**							
														***							
												3	1	***							
														***							
														***							
														***							
												4a	1	***							
												4b	1	**							
														***							
														***							

References

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies, 51*(1), 143-174.

Anonymous. (2015). The Case of the Hypothesis That Never Was; Uncovering the Deceptive Use of Post Hoc Hypotheses. *Journal of Management Inquiry, 24*(2), 214-216.

Artz, M., Homburg, C., & Rajab, T. (2012). Performance-measurement system design and functional strategic decision influence: The role of performance-measure properties. *Accounting, organizations and society, 37*(7), 445-460.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility: Survey sheds light on the 'crisis' rocking research. *Nature, 533*(7604), 452–454.

Ballas, A., & Theoharakis, V. (2003). Exploring diversity in accounting through faculty journal perceptions. *Contemporary Accounting Research, 20*(4), 619-644.

Bamber, L. S., Christensen, T. E., & Gaver, K. M. (2000). Do we really 'know'what we think we know? A case study of seminal research and its subsequent overgeneralization. *Accounting, organizations and society, 25*(2), 103-129.

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. (2016). Questions about questionable research practices in the field of management a guest commentary. *Journal of Management, 42*(1), 5-20.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323-338.

Bedeian, A. G. (2004). Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning & Education, 3*(2), 198-216.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education, 9*(4), 715-725.

Bedford, D. S. (2015). Management control systems across different modes of innovation: Implications for firm performance. *Management Accounting Research, 28*, 12-30.

Bedford, D. S., & Malmi, T. (2015). Configurations of control: An exploratory analysis. *Management Accounting Research, 27*, 2-26.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533.

Benjamini, Y., & Hechtlinger, Y. (2014). Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. *Biostatistics, 15*(1), 13-16.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization, 15*(3), 423-436.

Bisbe, J., & Malagueño, R. (2012). Using strategic performance measurement systems for strategy formulation: Does it work in dynamic environments? *Management Accounting Research, 23*(4), 296-311.

Bonner, S. E., Hesford, J. W., Van der Stede, W. A., & Young, S. M. (2006). The most influential journals in academic accounting. *Accounting, organizations and society, 31*(7), 663-685.

Borkowski, S. C., Welsh, M. J., & Zhang, Q. M. (2001). An analysis of statistical power in behavioral accounting research. *Behavioral Research in Accounting, 13*(1), 63-84.

Butler, N., Delaney, H., & Spoelstra, S. (2017). The gray zone: Questionable research practices in the business school. *Academy of Management Learning & Education, 16*(1), 94-109.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Chan, T. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436.

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67*(6), 691-700.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61*(4), 287-292.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered reports at AIMS neuroscience and beyond. *AIMS Neuroscience, 1*(1), 4-17.

Chen, W., & Tan, H.-T. (2013). Judgment effects of familiarity with an analyst's name. *Accounting, organizations and society, 38*(3), 214-227.

Chua, W. F. (1986). Radical developments in accounting thought. *Accounting review, 61*(4), 601-632.

Dyckman, T. R., & Zeff, S. A. (2014). Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons, 28*(3), 695-712.

Evans, J. H., Feng, M., Hoffman, V. B., Moser, D. V., & Van der Stede, W. A. (2015). Points to Consider When Self-Assessing Your Empirical Accounting Research. *Contemporary Accounting Research, 32*(3), 1162-1192.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21*(5), 1180-1187.

Garud, R. (2015). Eyes wide shut? A commentary on the hypothesis that never was. *Journal of Management Inquiry, 24*(4), 450-454.

Gelman, A., & Loken, E. (2014a). Ethics and Statistics: The AAA Tranche of Subprime Science. *CHANCE, 27*(1), 51-56.

Gelman, A., & Loken, E. (2014b). The statistical crisis in science. *American Scientist, 102*(6), 460.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science the idol of a universal method for scientific inference. *Journal of Management, 41*(2), 421-440.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine, 8*(341), 341ps312.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological bulletin, 82*(1), 1.

Guerreiro, M. S., Rodrigues, L. L., & Craig, R. (2012). Voluntary adoption of International Financial Reporting Standards by large unlisted companies in Portugal–Institutional logics and strategic responses. *Accounting, organizations and society, 37*(7), 482-499.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods, 12*(3), 179-185.

Hardwicke, T. E., Jameel, L., Jones, M., Walczak, E. J., & Weinberg, L. M. (2014). Only Human: Scientists, Systems, and Suspect Statistics. *Opticon1826, 25*(16), 1-12.

Henri, J.-F., & Journeault, M. (2010). Eco-control: The influence of management control systems on environmental and economic performance. *Accounting, organizations and society, 35*(1), 63-80.

Ho, J. L., Wu, A., & Wu, S. Y. (2014). Performance measures, consensus on strategy implementation, and performance: Evidence from the operational-level of organizations. *Accounting, organizations and society, 39*(1), 38-58.

Hubbard, R., & Lindsay, R. M. (2013). The significant difference paradigm promotes bad science. *Journal of Business Research, 66*(9), 1393-1397.

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals* (3. ed., repr. / with text rev. and notes by P.H. Nidditch ed.). Oxford: Clarendon.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648.

Ioannidis, J. P. (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics, 15*(1), 28-36.

Ioannidis, J. P., & Doucouliagos, C. (2013). What's to know about the credibility of empirical economics? *Journal of Economic Surveys, 27*(5), 997-1004.

Ittner, C. D. (2014). Strengthening causal inferences in positivist field studies. *Accounting, organizations and society, 39*(7), 545-549.

Jager, L. R., & Leek, J. T. (2013). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics, 15*(1), 1-12.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again.... *Science, 334*(6060), 1225.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science, 23*(5), 524-532.

Kane, E. J. (1984). Why journal editors should encourage the replication of applied econometric research. *Quarterly Journal of Business and Economics, 23*(1), 3-8.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience, 12*(5), 535-540.

Lachmann, M., Trapp, I., & Trapp, R. (2017). Diversity and validity in positivist management accounting research—A longitudinal perspective over four decades. *Management Accounting Research, 34*, 42-58.

Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review, 7*(3), 471-479.

Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research, 11*(1), 33.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585.

Luft, J., & Shields, M. D. (2014). Subjectivity in developing and validating causal explanations in positivist accounting research. *Accounting, organizations and society, 39*(7), 550-558.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological bulletin, 70*(3p1), 151-159.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods, 1*(2), 130.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research, 1*(2), 161-175.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review, 104*(1), 277-290.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737-738.

Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. (2015). Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013. *Communication Methods and Measures, 9*(4), 193-207.

McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regressions. *journal of Economic Literature, 34*(1), 97-114.

McNutt, M. (2014). Raising the bar. *Science, 345*(6192), 9-9.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology, 46*(4), 806.

Merchant, K. A. (2010). Paradigms in accounting research: A view from North America. *Management Accounting Research, 21*(2), 116-120.

Merton, R. K. (1942). Note on science and democracy. *Journal of Legal & Political Sociology, 1*, 115.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago and London: University of Chicago press.

Moonesinghe, R., Khoury, M. J., & Janssens, C. J. (2007). Most published research findings are false--but a little replication goes a long way. *PLoS medicine, 4*(2).

Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *British journal of pharmacology, 172*(8), 2126-2132.

Munafò, M., Noble, S., Browne, W. J., Brunner, D., Button, K., Ferreira, J., . . . Lindquist, M. (2014). Scientific rigor and the art of motorcycle maintenance. *Nature biotechnology, 32*(9), 871-873.

Nosek, B., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., . . . Christensen, G. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348*(6242), 1422-1425.

Nosek, B., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631.

Nuzzo, R. (2014). Statistical errors. *Nature, 506*(7487), 150-152.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. *Journal of Management, 43*(2), 376-399.

Ohlson, J. A. (2015). Accounting research and common sense. *Abacus, 51*(4), 525-535.

Perreault, S., & Kida, T. (2011). The relative effectiveness of persuasion tactics in auditor–client negotiations. *Accounting, organizations and society, 36*(8), 534-547.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov, 10*(9), 712-712.

Shields, M. D. (1997). Research in management accounting by North Americans in the 1990s. *Journal of Management Accounting Research, 9*, 3.

Siegfried, T. (2010). Odds Are, It's Wrong: Science fails to face the shortcomings of statistics. *ScienceNews, 177*(7). Retrieved from https://www.sciencenews.org/article/odds-are-its-wrong

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature, 526*(7572), 189-191.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534.

Starbuck, W. H. (2016). 60th Anniversary Essay How Journals Could Improve Research Practices in Social Science. *Administrative Science Quarterly*, 1-19.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association, 54*(285), 30-34.

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the P value is not enough. *Journal of graduate medical education, 4*(3), 279-282.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101.

Tullock, G. (2001). A comment on Daniel Klein's" a plea to economists who favor liberty". *Eastern Economic Journal, 27*(2), 203-207.

Van der Stede, W. A., Young, S. M., & Chen, C. X. (2005). Assessing the quality of evidence in empirical management accounting research: The case of survey studies. *Accounting, organizations and society, 30*(7), 655-684.

Van Noorden, R. (2014). Science joins push to screen statistics in papers: New policy follows efforts by other journals to bolster standards of data analysis. Retrieved from http://www.nature.com/news/science-joins-push-to-screen-statistics-in-papers-1.15509

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-values: context process, and purpose. *The American Statistician, 70*(2), 129-133.

Yong, E. (2012). Replication studies: Bad copy. *Nature, 485*(7398), 298-300.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Med, 5*(10), e201.

Ziliak, S. T. (2016). Statistical significance and scientific misconduct: Improving the style of published research papers. *Review of Social Economy, 74*(1), 83-97.

Ziliak, S. T., & McCloskey, D. N. (2004a). Significance redux. *The Journal of Socio-Economics, 33*(5), 665-675.

Ziliak, S. T., & McCloskey, D. N. (2004b). Size matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics, 33*(5), 527-546.

Chapter 4

A STUDY ON THE CRITERIA OF INTERNAL TRANSPARENCY, EFFICIENCY AND EFFECTIVENESS IN MEASURING LOCAL GOVERNMENT PERFORMANCE

Author: Kristian Mohr Røge and Niels Joseph Lennon

Abstract Public managers perceives performance measurement as an indispensable element in modernising the public sector in achieving 'more for less', despite research evidencing that performance measurement in the public sector is risky in terms of unexpected and undesirable effects. Through a case study of the introduction of a performance measurement system (PMS) in a Danish municipality, our analysis focuses on how internal transparency unfolds in terms of addressing and aligning performance measures with the criteria of efficiency and effectivity. We conclude that there are two main reasons why the PMS fails to address the desired effects of modernising the municipality. First, an overarching focus by top management on outcome measurement results in an inattention on measuring resource consumption and efficiencies, and second, an empowerment of lower level managers in the formulation of KPIs results in competence problems related to PMS design, which results in a large part of the KPIs being milestone measures and designed to constrain the influence of the PMS on daily activities. These matters hinder the organisation's ability to design KPIs that attribute costs to outputs, and therefore render the ambition of internal transparency impossible to achieve for the municipality. The consequence is a PMS that loses track of its very purpose and for top management it provides a false sense of security in the optimisation of scarce resources.

Keywords: Performance measurement, transparency, efficiency, effectivity, public sector

1. Introduction

For a long time, the public sector has been exposed to allegations of wastefulness and inefficiency (Modell, 2005). As a consequence, *'value for money'* has become an important, nontrivial aspect of local government management (Arnaboldi, Lapsley, & Steccolini, 2015; Kloot & Martin, 2000), and performance measurement is now considered an indispensable element in modernizing local government entities (Bouckaert & Peters, 2002; Cuganesan, Guthrie, & Vranic, 2014; Fryer, Antony, & Ogden, 2009).

However, public sector performance measurement is not a new concept (Ridley & Simon, 1938; Simon, 1937). Since Simon's seminal work (1937), performance measurement has been considered a tool enabling public managers to answer the following questions: (1) *'how adequate and effective is our service performance?',* and (2) *'how efficient are we in providing these services?'* (Simon, 1937). Thus, public managers are interested in the factors that render public sector organisations efficient or inefficient as well as in the measures that inform improvement of public sector efficiency.

The role of performance measurement in improving accountability, decision-making, and ultimately, public sector performance is often taken for granted (OECD, 1994, 1997; Van Thiel & Leeuw, 2002). However, the literature on performance measurement in the public sector is loaded with contesting opinions on its applicability (Johnsen & Vakkuri, 2006; Modell, 2005; Poister & Streib, 1999). Empirical results have evidenced that implementation of performance measurement systems (PMS) in the public sector has not resulted in the expected improvements in performance, accountability, transparency and quality of services (Arnaboldi et al., 2015; Fryer et al., 2009; Hood & Dixon, 2015).

It has been argued that the inadequacy of performance measurement is due to an unresolved issue with formulating performance measures (Bouckaert & Peters, 2002; Fryer et al., 2009; Lapsley, 2009; Van Thiel & Leeuw, 2002). This problem is attributed to the lack of a single, satisfactory, overall measure of performance comparable to the measurement of the financial performance of private organisations (Anthony & Young, 1999) along with the intangible nature of public services. For public organisations, it is easy to measure how much work has been done – but not on how well it was done, nor whether the particular work undertaken was appropriate to the desired end (Ridley & Simon, 1937). As financial measures are unsuitable for measuring the performance of public organisations, Ridley and Simon (1938) developed other criteria for the appraisal of public organisations, namely that performance measurement should measure the attainment of objectives and the efficiency in doing so. These concerns have later translated into the performance measurement criteria of efficiency and effectiveness (Anthony, 1965).

We aim at exploring how the efficiency and effectiveness criteria relate to the inadequacy of PMS implementations in public sector organisations; we argue that a measurement of these two criteria must be central for public sector PMS to achieve the strategic objectives with efficient

resource consumption under financial constraints. The efficiency and effectiveness criteria render the resource flow from costs, through outputs, to outcome transparent and manageable (Anthony, 1965; Hood, 1996; Vigoda-Gadot & Meiri, 2008). If a PMS fails to provide internal transparency on the resource consumption of output and outcome, it would in theory, be unable to inform decision making about improvements in organisational performance (Kaplan, 2001; Ridley & Simon, 1938; Simon, 1937). To explore this, we study the implementation of a PMS in a Danish municipality from the first working document to the latest iteration. We focus on the construction of KPIs and their ability to facilitate improvements in organisational performance.

The paper contributes to the theory and practice of public sector performance measurement, with insight into the importance of the role of design criteria efficiency and effectiveness in creating a functioning and successful PMS. We illustrate that if the two criteria are not met, a PMS cannot create internal transparency on the relationship between resource consumption and results. We therefore conclude that due to the particular formulation of KPIs, this PMS implementation loses track of its very purpose, namely to direct actions and decisions towards an efficient and effective use of resources in accomplishment of the municipality's strategic objectives. The PMS therefore creates a false sense of security for the optimization of scarce resources.

2. Theoretical framework

The public sector is not operated for profit (Anthony & Young, 1999), and the techniques of cost accounting only have limited applicability here. Other criteria for appraising local government activities therefore need to be developed (Ridley & Simon, 1938).

With the introduction of performance measurement, the question of what to measure arose. It is not enough to measure expenditure, like if a shopper for example told you that *'I am a very efficient shopper. I only spent five dollars today'.* The response would be *'that is all very well, but what did you get for your five dollars?'* (Ridley & Simon, 1937). On the other hand, it is not enough merely to measure effort, as for example when measuring services in *'man-hours'*; a reference is required to assess whether it was too little or too much. It is easy to make a measure of performance, that is, the effect of applying effort. For example, in the police force, a performance measure could be the number of criminals apprehended, and in public education it could be the number of students educated (Ridley & Simon, 1938, 1943). But measures as these, however useful they might seem, are inadequate for performance measurement purposes. They inform us on the extent of work done, but we also need to know the quality of the work done, and whether the particular work undertaken was appropriate to the desired end. This leads to the following definitions of effectiveness *"a measurement of the result of an effort or performance indicates the effect of that effort or performance in accomplishing its objective"* (Ridley & Simon, 1938, p. 21) or *"Effectiveness relates to the accomplishment of the coöperative*

purpose... When a specific desired end is attained we shall say that the action is 'effective'" (Barnard, 1938, p. 60).

These definitions show that effectiveness is a measure of the degree of the attainment of results, but it leaves one important question unanswered, namely how efficient management is in attaining the objective. This is of particular importance to public managers due to the limited public resources both in capital and human terms. Therefore, a public manager must maximise the attainment of organisational objectives through efficient and effective employment of the limited resources available (Kaplan, 2001; Modell, 2005; Ridley & Simon, 1938). This leads us to the following two definitions of efficiency: The first definition is efficiency as *"the optimum relationship between input and output"* (Anthony, 1965, p. 28), which slightly differs from the earlier definition by Ridley and Simon (1938, p. 23) *"the efficiency of administration is measured by the ratio of the effects actually obtained with the available resource to the maximum effects possible with the available resources"*. These definitions imply that the more units of output obtained from a given input, the more efficient the process.

Based on these definitions, which are central to PMS theory, a PMS is defined as *"a set of metrics used to quantify both efficiency and effectiveness of actions"* (Neely, Gregory, & Platts, 1995, p. 81) and a KPI is *"a metric used to quantify the efficiency and/or effectivity of an action"* (Neely et al., 1995, p. 80). To measure efficiency and effectiveness, four types of KPIs exist: input, process, output and outcome (Ammons, 1995; Hoque & Adams, 2011; Pollanen, 2005). Input measures are a quantification of resources used in providing a service. Output measures indicate the amount of work completed. Process measures are the ratio between input and output, i.e. efficiency. Outcome measures indicate the effect of services provided, and if the objective is attained, we can say that the effort was effective.

At a first glance, the concepts of efficiency and effectiveness seem contradictory, but management control can facilitate that resources are obtained and used efficiently and effectively in the accomplishment of the organisation's objectives (Anthony, 1965). It is the purpose of managers to balance them. To exemplify, a business unit that performs a service with the lowest possible resource consumption, may be highly efficient, but if its services fail to contribute adequately to the achievement of the organisation's goals, it is not deemed effective (Anthony & Govindarajan, 2003). Thus, without efficiency, knowledge on the resource consumption of an achieved outcome would not be available; while on the other hand, without effectiveness, it would not be known whether the output leads to the achievement of the objective. Performance measurement enables internal transparency[1] allowing efficiency and effectiveness improvements to take place in an 'optimised' manner (Ahrens & Chapman, 2004; Lapsley & Ríos, 2015).

The heart of any PMS is therefore its ability to balance efficiency and effectiveness (Pollanen, 2005), and only if these two criteria are covered satisfactorily, does it produce internal transparency on organisational performance. When internal transparency is created, resource

consumption and strategic effects are explicitly linked rendering efficiency and effectiveness improvements possible. A PMS that provides internal transparency is a system that balances efficient and effective resource consumption by directing actions and decisions toward the achievement of strategic objectives (Anthony & Govindarajan, 2003; Lapsley & Ríos, 2015; Pollanen, 2005).

However, before efficiency and effectiveness can be measured, the objectives of the organisation must be defined. For the public sector, the task of defining quantifiable objectives constitutes one of the most difficult tasks in the entire field of measurement (Ridley & Simon, 1938), because what is the role of the public sector and what is good performance? (Fryer et al., 2009; Van de Walle, 2008). Conventional accounting statements do not reflect the performance of public organisations, as healthy finances contain no information on the result of services (Kaplan, 2001). In addition, local government entities rarely have objectives as clearly defined and generally accepted as those of, for instance, a fire department or a police station. Aims such as *'improve health'*, *'develop good citizens'*, and *'high quality education'* must be stated in much more tangible and objective terms before they can be adapted to performance measurement (Ridley & Simon, 1938, 1943). Often, it is no easy task for public managers to formulate measureable objectives, as *"non-profit organisations frequently have goals that are amorphous and offer services that are intangible"* (Forbes, 1998, p. 184).

Therefore, it is of interest to investigate the interplay of these criteria for the functioning of a PMS, as an inaccurate measurement of performance may give managers a false sense of security about reaching their objectives while misdirecting resources and activities. In such a scenario, the PMS would potentially be counterproductive to the optimization of the finite resources (Bouckaert & Peters, 2002), as the PMS would be unable to support the process of management control or the strategic planning on taking 'intelligent decisions' that lead to performance improvements (Anthony, 1965).

On this backdrop, we express our research aim more specifically. Using internal documents and interviews, we analyse how the criteria of efficiency and effectiveness influence the functioning of a PMS in a Danish municipality. In particular, we will address the following research question: To what extent has the management of the municipality met the criteria of efficiency and effectiveness in their performance contracts between top management and organisational units, and what role does the attainment of the criteria of measurement play in ensuring a functioning PMS?

3. Research method

The paper reports a qualitative field study conducted in a large Danish municipality. The municipality was formed in 2007 in connection with a large local government reform where the number of Danish municipalities was reduced from 271 to 98. The reform was orchestrated by

a commission of experts who recommended that the overall size of Danish municipalities be significantly increased to accommodate the requirements of a modern society. The commission recommended that the size of municipalities be at least 30,000 citizens or otherwise engage in legally binding cooperation with larger municipalities. After the merger, the size of the case study municipality was approximately 60,000 citizens in 2016, with a budgeted income of DKK 3.5 billion and the municipality employs around 3,600 people. The merger called for the development of a new joint PMS.

We gained access to the municipality through 'cold calling', but we initially reached out to several Danish municipalities as we needed to identify a municipality with a deep engagement in performance measurement. The rationale for the case selection was to find an illustrative case that could provide a rich description of a 'real-world' situation (Flyvbjerg, 2011; Otley & Berry, 1994; Stake, 1994) where performance measurement was utilized in trying to assure that resources were used efficiently and effectively in the accomplishment of strategic objectives. By studying this municipality, we follow a recent suggestion in management accounting research, namely to focus on the technical core of the subject and to conceptualise the empirical findings into something that can develop and support practice (Baldvinsdottir, Mitchell, & Nørreklit, 2010; Van Helden & Northcott, 2010).

Our primary data consists of the performance contracts and strategic documents covering (the lifetime of) the PMS from 2007 to 2016. To increase the reliability and validity of our interpretation of the documents through data triangulation, we conducted qualitative interviews (Qu & Dumay, 2011; Yin, 2015) and collected secondary data, such as internal documents including measurement guidelines and internal surveys on the PMS (Ahrens & Chapman, 2006; Scapens, 2004; Vaivio, 2008).

The interviews served the purpose of deepening our understanding of how and why particular KPIs were formulated. To this end, we used semi-structured interview guides to lead the interviews, which were based on a preliminary analysis of the performance contracts. One of the authors conducted the transcription of the interviews, and NVivo was used to organize and code the data. The analysis strategy was a 'thematic analysis' (Guest, MacQueen, & Namey, 2011; Saldaña, 2015) where main themes were identified and coded.

The interviewees consisted of three key informants, an executive director, a support unit manager and a consultant; they were selected due to their privileged insight as they were in charge of the implementation of the PMS in 2007. The rest of the interviewees, a business unit manager and two sub-business unit managers were selected based on their hierarchal position in the business unit under investigation. In total, eight semi-structured interviews, lasting between 40 and 90 minutes, were conducted, transcribed, and analysed. The interview format was adjusted to the level of seniority and the respondent's area of responsibility. The fieldwork took place in the spring of 2014 and was followed up in the summer of 2016.

4. The municipality – contextual background

In 2007, top management and the city council decided to develop a new PMS which was to be constructed as a set of performance contracts between all hierarchal levels of the municipality. The PMS was intended to support strategic decision-making and management control as well as to provide all levels of management with information for facilitating an optimization of scarce resources taking service quality into account. The theoretical inspiration for the PMS, which the municipality developed, was the Balanced Scorecard (BSC); this became their design baseline. Using the BSC as their baseline prompted by Local Government Denmark[2]'s recommendation (Kræmmergaard, Rikhardsson, & Nielsen, 2006; Lauritsen & Sprong, 1999) and that the consultant in charge of implementing the new PMS also advocating for the BSC.

The design of the PMS

When developing the PMS in 2007, it was designed with the four perspectives 'human resources', 'physical and financial resources', 'processes', and 'customer satisfaction and effects'. However, in 2011, the PMS went through major modifications and the four perspectives were reduced to three. The modification resulted in the removal of the financial perspective from the PMS. This was due to a situation where resource consumption all too often was equalled with service quality. For example, quality in teaching was equalled by cost per student. Top management intended to hinder the use of resource consumption as a quality indicator instead directing the KPIs towards measuring output and outcome. Top management was aware of the risk of separating the PMS from cost information, which could disconnect it from budget decisions. However, they believed it was a necessary change, and it did not mean that financial measures were to be excluded; the financial measures could still be measured within the remaining three perspectives.

The three remaining perspectives were defined as follows: the 'human resource' perspective concerned the level of human capabilities needed for steering the future development of the municipality in achieving its strategic objectives. This perspective mainly consisted of KPIs that measure the concepts of 'human capital', 'values' and 'motivation'. The 'process' perspective focused on the internal conditions for the municipality to succeed in fulfilling the demand of their 'customers' and, in addition, a focus on casework time, organisational structure, organisational culture, corporate values, control tools and the management foundation was emphasised here. The first two perspectives focused on the measurement of resource consumption and efficiency. The 'effect' perspective visualised the link between output, outcome, and the strategic objectives and is where the effectiveness of actions was measured. A measurement of, for example, satisfaction levels and service outcomes was part of this perspective.

The municipality's 2011-2014 strategy consists of two strategic themes: *'optimal resource utilization and high quality' and 'growth and development'*; and to accomplish the two themes,

each business unit should formulate a set of strategic objectives. To accomplish the strategic objectives within the themes, each business unit within the municipality was required to engage in a performance contract with the executive board; this contract had to contain a set of KPIs, targets and initiatives.

To support the formulation of KPIs, the Knowledge and Strategy support unit developed comprehensive material to guide the managers along with providing mandatory courses. The guidelines were developed to ensure 'systematic', 'consistent' and 'credible' KPIs. Knowledge and Strategy also formulated seven principles that should always be taken into consideration when formulating a KPI. Consequently, a KPI should always be (1) measureable, (2) unambiguous, (3) communicable, (4) accepted by core stakeholders, (5) realistic, (6) applicable, but still ambitious and challenging, and (7) time based, such that time of delivery is known to all.

5. Analysis of the performance contract KPIs

This section analyses to which extent the performance contract KPIs conform to the performance measurement criteria of efficiency and effectiveness. The analysis focuses on the strategic theme *of 'optimal resource utilization and high quality'* and related objectives, as this theme represents the closest link to the original purpose of performance measurement and management control (Anthony, 1965; Ridley & Simon 1938). To identify whether a balanced measurement of efficiency and effectiveness are in place in the performance contracts, this is the place to look. We analyse the KPIs in the performance contract under the strategic objectives of *'healthy finances and high work quality', 'collaborate control, efficiency improvements and digitalization'*, and *'innovative capacity'.*

The decision to analyse the performance contract set up between the executive board and the business unit, 'Land, City and Culture', was made in consultation with the support unit manager responsible for the PMS. He believed that this business unit would be representative for the implementation and use of the PMS, and the unit had been very co-operative in regard to the PMS. From a methodical point of view, we are able to obtain a deeper understanding of the business unit and its performance contract by focusing on a single performance contract rather than broadening the study to cover more than one. Appendix A includes an organogram of the business unit. This provides some clarity to what their operational tasks are.

In the following, we present our analysis of the performance contract; we only present the part of the contract that contains KPIs and targets; the rest of the contract is a long list of initiatives irrelevant to the scope of this paper. The human resource perspective is only part of the executive board's performance contract and is therefore not part of the analysis.

The process perspective

The process perspective is confined by the two objectives *'collaborate control, effectiveness improvements and digitalisation' and 'innovative capacity'*. Figure 1 presents an

overview of the KPIs formulated to ensure the accomplishment of the two strategic objectives. The process perspective contains five KPIs, which ideally are intended to measure input and processes. This part of the performance contracts is focused on answering 'how efficient are we in providing our services?'

Perspective:	**Process**		
Theme:	**Optimal resource utilization and high quality**		
Objective:	**Collaborate control, effectiveness and digitalisation**		
Measure	**Goal**	**Type of measure**	**Unit of measure**
Cross-disciplinary collaboration	Maximize	Quantitative	1 decimal
Description:	The following question in the job satisfaction survey (2012): "I feel that the cross-disciplinary collaboration crosswise agreement units and business units", must achieve a score of equal or above 3.8 in the fourth quarter of 2014. Measured as an internal investigation.		
Objective:	**Innovative capacity**		
Measure	**Goal**	**Type of measure**	**Unit of measure**
# completed innovation courses		Qualitative	Achieved Yes/No
Description:	A minimum of one yearly innovation course is achieved in all agreement units		
Citizen and company participation		Qualitative	Achieved Yes/No?
Description:	Two voluntary workshops/events to be conducted in the period 2013-2014.		
International cooperation		Qualitative	Achieved Yes/No
Description:	Participating in at least 5 international collaborations across the business unit Land, City and Culture in the period 2013 – 2014.		
# implemented ideas from the innovation courses	Maximise	Quantitative	1 decimal
Description:	Number of implemented ideas from the innovation course. Target: at least one idea is implemented in all agreement units each year.		

Fig. 1. translated extract from the performance contract of the business unit Land, City and Culture

The first strategic objective consists of only one KPI, which measures cross-disciplinary collaboration between business units and agreement units[3]. This is achieved through a job satisfaction survey including a question on the satisfaction level with cross-collaboration, which is measured on a scale from one to five. The second objective, *'innovative capacity'*, is composed of four KPIs, which focus on innovation courses, idea implementation, and citizen, company and international cooperation. The four KPIs, respectively, measure the number of completed innovation courses, number of ideas implemented from these courses, the number of voluntary workshops held with citizens and companies, and lastly, the number of international collaborations across the business unit.

According to the executive strategy plan, the process perspective should include measures of input and process on services provided. However, none of the five KPIs measures any aspect of resource consumption or efficiency in attaining the strategic objectives. Instead, the KPIs measure the progress of initiatives, which is Kaplan (2001) categorises as milestone

measurement. Such measures convey no information on the effect of the application of effort [effectiveness] or the accomplishment relative to the resources used [efficiency], and as such, they cannot be classified as performance metrics and should instead be classified as *only* measures (Ijiri, 1975).

A lack of input and process KPIs is of concern if the PMS must facilitate managers in optimising the relationship between input and output (Anthony, 1965). These KPIs therefore conflict with the criteria for measuring performance (Neely et al., 1995; Ridley & Simon, 1938) and with providing a foundation for management control or strategic planning (Anthony, 1965). The PMS is unlikely to be able to facilitate an efficient use of the limited resources available when these resources are not measured. However, it is of course possible that the fulfilment of the targets of the KPIs and initiatives could lead to efficiency improvements; however, it is not because of the PMS creating internal transparency which renders the link between resource consumption and outputs explicit so that it can be optimised.

The 'effect' perspective

The effect perspective is constituted by the strategic objective of *'healthy finances and high work quality'*; figure 2 presents an overview of the related KPIs. The effect perspective contains seven KPIs which ideally are intended to measure outputs and outcomes. This part of the performance contract is focused on answering 'how adequately and effective are we in performing our services?'

Perspective:	**Effect**		
Theme:	**Optimal resource utilization and high quality**		
Objective:	**Healthy finances and high work quality**		
Measure	**Goal**	**Type of measure**	**Unit of measure**
Financial culture	Maximise	Quantitative	1 decimal
Description:	- To gain external funding of 10 million Danish kroner yearly.		
Financial culture	Maximise	Quantitative	1 decimal
Description:	Achieve 10% reduction in energy expenses yearly (in comparison with invested).		
Financial culture		Qualitative	Achieved Yes/No
Description:	Late 2013, three scenarios for alternative ways of operating are described with corresponding unit costs and the financial potential.		
Case work time	Minimize	Quantitative	1 decimal
Description:	Total casework time for the approval of livestock is less than 7 months.		
Case work time	Minimize	Quantitative	1 decimal
Description:	Total casework time for the approval of building projects is less or equal to 28 days.		
Upheld verdicts	Maximize	Quantitative	1 decimal
Description:	90% of the complaints against administrative verdicts are upheld in court		
Appropriate service level	Maximize	Quantitative	1 decimal
Description:	75% of the respondents must answer "satisfactory" or better.		

Fig. 2. translated extract from the performance contract of the business unit Land, City and Culture

The KPIs are divided into four measurement subgroups, (1) financial culture, (2) casework time, (3) upheld verdicts, and (4) appropriate service level. The first three KPIs comprise the measurement of financial culture, aimed at achieving annual external funding of 10 million DKK, a 10-percent reduction in energy expenses, and developing three scenarios for alternative ways of operating. The two KPIs from group two measure total building licence casework time (target is less than 7 months) and the approval of livestock (target less than 28 days). The last two KPIs measure whether administrative decisions are upheld in court when contested and the satisfaction level of the services provided through surveys can both be considered outcome measures.

According to the executive strategy plan, the effect perspective should measure the effectiveness of the services provided by the municipality. Output measures are not interesting in themselves as they contain no referent to assess whether it was too little or too much or at a wrong quality level, which is why this perspective should rather focus on the 'effect of the application of effort' (Ridley & Simon, 1938). Managers need to know how well a particular piece of work was done and whether it was appropriate to the desired end. The KPIs of the effect perspective partly succeed in doing this, but there is no reference to the use of resources. For example, the external funding KPI measures funding received, but disregards the costs of applying for external funding or the potential risk of failing in attaining the resources. Another example is the KPIs measuring total casework time on the approval of building projects or livestock. These KPIs only measure the output on these services, but disregards if the lowering of casework time is a consequence of changes in allocation of financial resources or 'man-hours'. As a result, the effectiveness of the services could be increasing, while efficiency is unknowingly decreasing. This problem is a common denominator for the KPIs in the effect perspective, i.e. that changes in effectiveness might be the result of an increase in resource consumption; this is contrary to the principle of optimising the balance between these two criteria. In consequence, the PMS cannot create internal transparency because it does not provide transparency on efficiency and effectiveness so that management can optimise the balance. As a result, there is a risk that the PMS is counterproductive to performance.

To summarise, the analysis evidences that the PMS partly answers the question of 'how adequately and effectively does it perform it services?', but it completely fails to answer, 'how efficiently are we in providing these services?' (Simon, 1937). It is therefore unlikely that the information produced by the PMS is useful to the municipality management in fulfilling their function of maximising the attainment of organisational objectives through efficient use of the limited resources at their disposal. As a consequence, top management does not know whether Land, City and Culture spend their resources in an optimal manner.

6. Discussion

The analysis confirms that to attain value from performance measurement, the choice of KPIs is one of the most critical challenges facing public managers (Ittner & Larcker, 1998). Top management not only removed the financial perspective in 2011, they were also strongly focused on the measurement of effectiveness and not resource consumption:

> *"Our main focus has not been on including measurements of cost... Instead, we focused on formulating 'good' KPIs that measured the effect of our initiatives and actions. Our focus therefore was on whether we achieved the desired results... It is difficult enough to create an organisation that can measure effectiveness.*
>
> (Support unit manager of Knowledge and Strategy)

Removing the financial perspective and emphasising effectiveness, top management disregarded any measurement of resource consumption.

The analysis also showed that Land, City and Culture were on the right track in their formulation of KPIs in the effect perspective, seeing that they measured results. However, we question whether these measured results are linked to the attainment of strategic objectives. In other words, their result measures did not necessarily measure effectiveness, because only when a specific desired end is attained can we claim that actions are 'effective' (Barnard, 1938). In this case, it is fair to question whether the fulfilment of the KPI targets leads to the accomplishment of the strategic objectives and themes when not one single KPI is measuring resource consumption or quality of service.

The challenge of formulating quantifiable objectives is known to be a difficult task for municipality managements (Ridley & Simon, 1938), as it is not easy to formulate objectives that are framed in a concrete form adaptable to performance measurement. The support unit manager responsible for the PMS is aware of this difficulty:

> *If it is impossible to formulate a proper KPI on outcome, so we have to descend a level in the measurement hierarchy and measure progress in initiatives instead.*
>
> (Support unit manager of Knowledge and Strategy)

This quote also explains why the PMS contained milestone measures. According to the manager, this happens when it is impossible to design outcome KPIs. However, as a result the KPIs are unsuited for management control and strategic planning (Anthony, 1965), i.e. the optimisation of scarce resources in the attainment of strategic objectives, the risk being that the PMS (at best) becomes an administrative cost and nothing more. At worst, it provides a false sense of security in the accomplishment of strategic objectives and misdirects resources and activities. It could become an Achilles' heel for the municipality in providing efficient and effective services (Bouckaert & Peters, 2002). For top management the formulation of KPIs is not satisfactory:

> *"There is a tendency among managers to select generic measures and targets, and in some cases, they let the data material decide which KPIs to formulate. This is highly unfortunate."*
>
> (Support unit manager of Knowledge and Strategy)

However, he does not question the lack of measuring and balancing efficiency and effectiveness. Instead, he highlights that there is an issue as regards formulating generic KPIs or letting the data material decide what to measure. That the municipality experiences issues with formulating KPIs represents a recurring and unsolved issue for the management of public organisations (Chang, 2007; Kasperskaya, 2008; Northcott & Ma'amora Taulapapa, 2012). The municipality tried to avoid this issue by developing material to guide the formulation of KPIs but it seems that they have failed.

What drove the formulation of KPIs for lower level managers?

Initially, the PMS was a top-down process where top management decided what to measure and how to measure it. This resulted in managers asking themselves if measures where formulated for the sake of measures? Top management realised the lack of ownership and commenced a long process of increasing sub-business unit managers' involvement and ownership in the formulation of KPIs:

> *"The way we now formulate KPIs is that we have a task force that collects information from our various sub-business units. The task force then carries the information into a staff meeting where we set up the performance contract... it is a dialogue-based process... It is, of course, important that the right KPIs are included. This can actually help us to keep on the right track...*
>
> (Business unit manager of Land, City and Culture)

The quote illustrates that lower level managers are included in the process of formulating KPIs. The business unit manager also considers the PMS a tool for ensuring that the unit is on the right track, i.e. accomplishes the strategic objectives, and emphasises that it is important to choose the 'right' KPIs. But this perception is not necessarily shared by the sub-business managers in Land, City and Culture:

> *"I feel that our business unit is very good at working together when formulating our performance contract, we (read: sub-unit managers) aren't steamrolled. We have the opportunity to influence the performance contracts when reaching the substance matter, and we can make suggestion for alterations. That is where we need to be smart and suggest what we know we can fulfil... It has to be realistic. Otherwise, it becomes fluffy. If we see that it is something that we are already working on, it makes more sense for us. We might broaden it a bit and it might receive more attention. For example, it could be a report that we intended to do anyway. Now it is just formulated as a KPI in the performance contract, as a multiple year target.*
>
> (Sub-business unit manager within Land, City and Culture)

It is worth noting that the sub-business unit managers feel they have leverage as regards the shape and targets of KPIs, which means that they are playing an important role in the

formulation of KPIs. However, another aspect of the quote is the expression that the KPIs are formulated as multiple year targets (milestone measurement) and that they should be 'smart' and suggest something they know is attainable. With influence come responsibility, but it would appear that the sub-business unit managers are not that concerned with the measurement of performance:

> *"The performance contracts are difficult when reaching the particularities.... the wording is not difficult, putting the right visions into words is not difficult, but it is difficult to formulate KPIs when reaching the substance... The performance contracts that are formulated now are actually formulated by the employees and lower level managers. This is how we initiate activities, which we want to do... For me, the performance contract is a bit like a memorandum about what we have to do this year"*
>
> (Sub-business unit manager within Land, City and Culture)

In addition, the sub-business unit manager argues that it is easy enough to formulate strategic objectives in the right words, i.e. 'optimal resource utilization and high quality', but it is difficult to measure the objectives. Defining quantifiable strategic objectives is perhaps one of the more difficult tasks of performance measurement in the public sector due to the intangible nature of their services (Forbes, 1998; Ridley & Simon, 1938). This quote evidences that this continues to cause problems for public managers despite it being discussed as early as the 1930s. It is also evidence of a major problem for performance measurement in the public sector, namely what is the role of the public sector and what is good performance? (Fryer et al., 2009; Van de Walle, 2008). It should be noted that the performance contract is formulated by lower level managers and employees, and they look on the performance contract as a memorandum of what they have to do each year and not as a measurement of performance. This perspective is irreconcilable with the purpose of performance measurement. When KPIs are formulated with this feature in mind, it is no wonder that the performance contract exhibits a large number of milestone measures thereby failing in creating internal transparency on performance.

The analysis evidenced that the criterion of efficiency was disregarded and that the outcome measurement was likely to be unrelated to the accomplishment of the strategic objectives, placing effectiveness in question. This results in a PMS that loses track of its very purpose, i.e. to direct actions and decisions towards an efficient accomplishment of strategic objectives. Consequently, the PMS does not create internal transparency, because no link between resource consumption and strategic effects is established. Therefore, it is impossible for managers to manage the relationship between efficiency and effectiveness. For managers engaging in management control and strategic planning, it is problematic that the information from the PMS is distorted or inaccurate. This renders the information unsuited for strategic decision-making, while also providing top management with a false sense of security in the optimisation of scarce resources. However, it should not be drawn from this study that the failure of performance

measurement is inevitable, but that an implementation of PMS should carefully consider the criteria for measuring performance, and that parts of public organisations are likely to be more suitable for measurement than others. Rarely does a municipality have equally tangible criteria for success as a fire department or police station. Aims, such as 'healthy finances', 'high work quality' or 'digitalization', must be stated in a much more tangible form before a 'proper' performance measurement is possible. We therefore argue for careful consideration of which areas to measure and how, instead of imposing PMS on the entire organisation. It is an ultimate premise for performance measurement that strategic objectives are formulated in a concrete and quantifiable way (Ridley & Simon, 1938).

To sum up the key points on why the PMS failed in answering the two questions of interest to public managers: (1) '*how adequately and effectively does it perform in its services?*' and, (2) *'how efficient are we in providing these services?'* (Simon, 1937). The analysis illustrates the following: First, top management removed the financial perspective from the PMS, which resulted in a reduced incentive for measuring resource consumption. Top management instead directed attention towards measuring output and outcome, while allowing for measures to be formulated as milestones, if results were deemed too difficult to quantify. Second, the PMS exhibited a disconnection between the outcome measurement and the strategic objectives thereby risking the effectivity of efforts. Third, decentralising the formulation of KPIs to lower level managers and employees resulted in a situation where lack of incentives and lack of PMS design competences had severe consequences for the formulation of KPIs, as the information provided by the PMS did not satisfy the proposed purpose of the system. Rather, the formulation of KPIs became severely influenced by data availability, choosing to measure what they knew they could fulfil, and a perception of the PMS as a memorandum. Fourth, lower level management also explained that they felt it was difficult to formulate KPIs that measured the strategic objectives formulated by top management.

7. Conclusion

To a large extent the PMS implementation process followed the suggestions for both private and public organisations (Fryer et al., 2009; Lilian Chan, 2004; Northcott & Ma'amora Taulapapa, 2012; Radnor & Lovell, 2003). Top management ensured a participative pre-implementation design process, the PMS had unlimited support from the executive board, adequate resources were made available to support initiatives, the perspectives of the PMS were selected with the intention to fit the contextual situation, and post-implementation reviews were conducted in 2008, 2012, 2013 and 2014 for all of management.

It was the ambition of top management that the PMS should facilitate efficiency and effectiveness improvements in the municipality's services. Notwithstanding the endeavours to develop a well-functioning and successful PMS, the analysis argues that the PMS fails in

accomplishing its purpose of direction, actions and activities toward the achievement of strategic objectives. It ends up being a PMS that is unable to create internal transparency, and therefore efficiency and effectiveness cannot be balanced through the activities of management control (Anthony, 1965). As a result, the KPIs fail in providing management with an explicit link between efficient resource consumption and effective realisation of the strategy. Yet, the PMS does provide a link between initiatives and strategic objectives, but the link is unmeasured in terms of performance changes and therefore the PMS cannot function as a foundation for management control and strategic planning. The PMS therefore ends up as an administrative burden and provides top management with a false sense of security in the optimisation of scarce resources.

Through this research, we partly explain the inadequacy of PMS implementations in the public sector, and we stress the importance for practitioners not to take the measurement of efficiency and effectiveness lightly. We argue that these criteria must be considered in the PMS design phase if a PMS is to be successful. If these criteria are not properly covered, the PMS cannot measure performance, and little does it then matter if the PMS is well received within the organisation. The biggest challenge for public sector management when implementing a PMS is therefore to ensure that they measure and balance efficiency and effectiveness; for this to be possible they have to formulate strategic objectives that are quantifiable in a concrete form as this is a basic requirement for any measurement.

It is unquestionable that there is a need for PMS to create a more efficient and effective public sector as resources is finite. Yet there is still no unanimity in research on what constitutes the best practice for implementing and creating a well-functioning PMS in the public sector. Efficiency and effectiveness are certainly two criteria that play a fundamental role in this accomplishment. This is a theoretical claim made already in the late 1930s by Ridley and Simon (1938) and Simon (1937) but seemingly forgotten or underappreciated in the performance measurement theories and practices of today.

NOTES

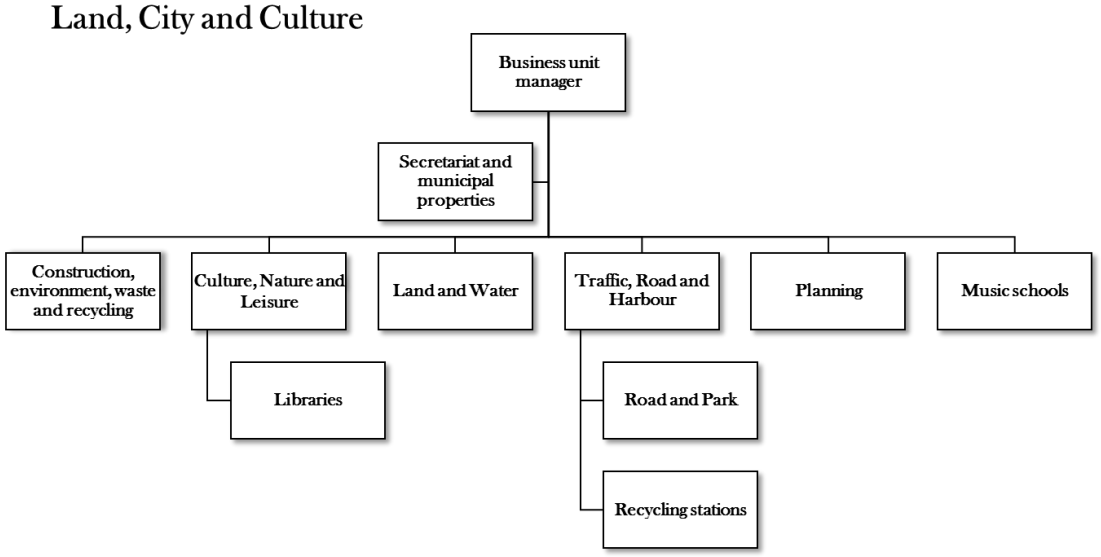
[1]Internal transparency is here defined as rendering the resource flow from costs to output to outcome transparent and manageable (Hood, 1996; Vigoda-Gadot & Meiri, 2008).

[2]Local Government Denmark is the association and interest organisation of the 98 Danish municipalities.

[3]Agreement units are organisational units that have their own budgets and management and are controlled trough contractual agreements. Examples of such units are schools, libraries, museums etc.

Appendix A

Organisational diagram for the business unit of Land, City and Culture



References

Ahrens, Thomas, & Chapman, Christopher S. (2004). Accounting for flexibility and efficiency: A field study of management control systems in a restaurant chain. *Contemporary Accounting Research, 21*(2), 271-301.

Ahrens, Thomas, & Chapman, Christopher S. (2006). Doing qualitative field research in management accounting: Positioning data to contribute to theory. *Accounting, Organizations and Society, 31*(8), 819-841.

Ammons, David N. (1995). Performance measurement in local government *Accountability for Performance: Measurement and Monitoring in Local Government* (pp. 15-32). Washington, DC: International City/County Management Association.

Anthony, Robert N. (1965). *Planning and Control Systems. A Framework for Analysis*. Boston: Graduate School of Business Administration, Harvard University.

Anthony, Robert N., & Govindarajan, Vijay. (2003). *Management control systems*. Boston: McGraw-Hill/Irwin.

Anthony, Robert N., & Young, David W. (1999). *Management control in nonprofit organizations*. Homewood, IL: Irwin/McGraw-Hill.

Arnaboldi, Michela, Lapsley, Irvine, & Steccolini, Ileana. (2015). Performance Management in the Public Sector: The Ultimate Challenge. *Financial Accountability & Management, 31*(1), 1-22.

Baldvinsdottir, Gudrun, Mitchell, Falconer, & Nørreklit, Hanne. (2010). Issues in the relationship between theory and practice in management accounting. *Management Accounting Research, 21*(2), 79-82.

Barnard, Chester Irving. (1938). *The functions of the executive*. Cambridge: Harvard university press.

Bouckaert, Geert, & Peters, B Guy. (2002). Performance measurement and management: The Achilles' heel in administrative modernization. *Public performance & management review, 25*(4), 359-362.

Chang, Li-cheng. (2007). The NHS performance assessment framework as a balanced scorecard approach: Limitations and implications. *International Journal of Public Sector Management, 20*(2), 101-117.
Cuganesan, Suresh, Guthrie, James, & Vranic, Vedran. (2014). The riskiness of public sector performance measurement: a review and research agenda. *Financial Accountability & Management, 30*(3), 279-302.
Flyvbjerg, Bent. (2011). Case Study. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage Handbook of Qualitative Research* (pp. 301-316). Thousand Oaks, CA: Sage.
Forbes, Daniel P. (1998). Measuring the unmeasurable: Empirical studies of nonprofit organization effectiveness from 1977 to 1997. *Nonprofit and Voluntary Sector Quarterly, 27*(2), 183-202.
Fryer, Karen, Antony, Jiju, & Ogden, Susan. (2009). Performance management in the public sector. *International Journal of Public Sector Management, 22*(6), 478-498.
Guest, Greg, MacQueen, Kathleen M, & Namey, Emily E. (2011). *Applied thematic analysis*. Thousand Oaks: Sage.
Hood, Christopher. (1996). Beyond "progressivism": a new "global paradigm" in public management? *International Journal of Public Administration, 19*(2), 151-177.
Hood, Christopher, & Dixon, Ruth. (2015). What We Have to Show for 30 Years of New Public Management: Higher Costs, More Complaints. *Governance, 28*(3), 265-267.
Hoque, Zahirul, & Adams, Carol. (2011). The rise and use of balanced scorecard measures in Australian government departments. *Financial Accountability & Management, 27*(3), 308-334.
Ijiri, Yuji. (1975). *Theory of accounting measurement*. Sarasota: American Accounting Association.
Ittner, Christopher D, & Larcker, David F. (1998). Innovations in Performance Measurement: Trends and Research. *10*, 205-238.
Johnsen, Åge, & Vakkuri, Jarmo. (2006). Is there a Nordic perspective on public sector performance measurement? *Financial Accountability & Management, 22*(3), 291-308.
Kaplan, Robert S. (2001). Strategic performance measurement and management in nonprofit organizations. *Nonprofit Management and Leadership, 11*(3), 353-370.
Kasperskaya, Yulia. (2008). Implementing the balanced scorecard: a comparative study of two Spanish city councils–an institutional perspective. *Financial Accountability & Management, 24*(4), 363-384.
Kloot, Louise, & Martin, John. (2000). Strategic performance management: A balanced approach to performance management issues in local government. *Management Accounting Research, 11*(2), 231-251.
Kræmmergaard, Pernille, Rikhardsson, Pall M, & Nielsen, Rune Ahlmann. (2006). *Økonomistyring i bevægelse-udfordringer og værktøjer*. København: Kommunernes Landsforening.
Lapsley, Irvine. (2009). New public management: The cruellest invention of the human spirit? *Abacus, 45*(1), 1-21.
Lapsley, Irvine, & Ríos, Ana-María. (2015). Making sense of government budgeting: an internal transparency perspective. *Qualitative Research in Accounting & Management, 12*(4), 377-394.
Lauritsen, Finn, & Sprong, Vibeke van der. (1999). *Balanced scorecard i kommunerne : helhed i ledelsens mål- og rammestyring*. København: Kommunernes Landsforening.
Lilian Chan, Yee-Ching. (2004). Performance measurement and adoption of balanced scorecards: a survey of municipal governments in the USA and Canada. *International Journal of Public Sector Management, 17*(3), 204-221.
Modell, Sven. (2005). Performance management in the public sector: past experiences, current practices and future challenges. *Australian Accounting Review, 15*(37), 56-66.
Neely, Andy, Gregory, Mike, & Platts, Ken. (1995). Performance measurement system design: a literature review and research agenda. *International Journal of Operations & Production Management, 15*(4), 80-116.

Northcott, Deryl, & Ma'amora Taulapapa, Tuivaiti. (2012). Using the balanced scorecard to manage performance in public sector organizations: Issues and challenges. *International Journal of Public Sector Management, 25*(3), 166-191.
OECD. (1994). *Performance management in government : performance measurement and results-oriented management*. Paris: Organisation for Economic Co-operation and Development.
OECD. (1997). *In search of results : performance management practices*. Paris: Organisation for Economic Co-operation and Development.
Otley, David T, & Berry, AJ. (1994). Case study research in management accounting and control. *Management Accounting Research, 5*(1), 45-65.
Poister, Theodore H, & Streib, Gregory. (1999). Performance measurement in municipal government: Assessing the state of the practice. *Public Administration Review, 59*(4), 325-335.
Pollanen, Raili M. (2005). Performance measurement in municipalities: Empirical evidence in Canadian context. *International Journal of Public Sector Management, 18*(1), 4-24.
Qu, Sandy Q, & Dumay, John. (2011). The qualitative research interview. *Qualitative Research in Accounting & Management, 8*(3), 238-264.
Radnor, Zoe, & Lovell, Bill. (2003). Success factors for implementation of the balanced scorecard in a NHS multi-agency setting. *International Journal of Health Care Quality Assurance, 16*(2), 99-108.
Ridley, Clarence E, & Simon, Herbert A. (1937). Technique of appraising standards. *Public Management, 19*(2), 46-49.
Ridley, Clarence E, & Simon, Herbert A. (1938). The criterion of efficiency. *The Annals of the American Academy of Political and Social Science, 199*(1), 20-25.
Ridley, Clarence E, & Simon, Herbert A. (1943). *Measuring municipal activities: A survey of suggested criteria for appraising administration*: The International City Managers' Association.
Saldaña, Johnny. (2015). *The coding manual for qualitative researchers*: Sage.
Scapens, Robert W. (2004). Doing case study research. *The real life guide to accounting research*, 257-279.
Simon, Herbert. (1937). Comparative statistics and the measurement of efficiency. *National Civic Review, 26*(11), 524-527.
Stake, RE. (1994). Case studies. InN. K. Denzin & YS Lincoln (Eds.), Handbook of qualitative research (pp. 236-247): Thousand Oaks, CA: Sage.
Vaivio, Juhani. (2008). Qualitative management accounting research: rationale, pitfalls and potential. *Qualitative Research in Accounting & Management, 5*(1), 64-86.
Van de Walle, Steven. (2008). Comparing the performance of national public sectors: conceptual problems. *International Journal of Productivity and Performance Management, 57*(4), 329-338.
Van Helden, G Jan, & Northcott, Deryl. (2010). Examining the practical relevance of public sector management accounting research. *Financial Accountability & Management, 26*(2), 213-240.
Van Thiel, Sandra, & Leeuw, Frans L. (2002). The performance paradox in the public sector. *Public Performance & Management Review, 25*(3), 267-281.
Vigoda-Gadot, Eran, & Meiri, Sagie. (2008). New public management values and person-organization fit: a socio-psychological approach and empirical examination among public sector personnel. *Public Administration, 86*(1), 111-131.
Yin, Robert K. (2015). *Qualitative research from start to finish*: Guilford Publications.

Chapter 5

THE ILLUSION OF 'OBJECTIVE AND RESULT-BASED MANAGEMENT': BEYOND AN NPM TOOL IN DENMARK

Author: Kristian Mohr Røge, Nikolaj Kure and Hanne Nørreklit

Abstract: In recent years, New Public Management (NPM) has been criticised for shaping a public sector that overemphasizes control and measurement at the expense of service quality and efficiency. Nevertheless, the Danish Ministry of Financial Affairs has accelerated the implementation of NPM by refining its standards and expanding its range of operation. A key element of this acceleration is the use of the NPM tool, 'objective and result-based management': a causal performance measurement system based on a contractual relationship between parties, typically a ministry and a governmental agency, such as a hospital or a university. This paper examines the conceptual qualities of the framework of 'objective and result-based management' used in the Danish public sector with a view to evaluate whether it may contain conceptual deficiencies that are predisposed for facilitating problems of effectiveness and efficiency. The paper found that the model was poorly conceptually outlined and with mismatches in the conceptual structure leading to a language game of illusions. The causal schematics outlined in the model are too vague, uncertain and general to guide actions that lead to the desired end of efficiency and service quality. It creates a social space where top managers are not made accountable and where employees are not disposed to develop effective practices. In conclusion, we find that the outlined performance management framework does facilitate the intended purpose of creating effective public sector institutions.

Keywords: NPM; Performance measurement; Causality; Validity; Public sector

1. Introduction

New Public Management (NPM) dates back to the early 1970s where politicians and economists became increasingly concerned about the growth in public spending and, in particular, the utility of this growth. Seeing that in general, private businesses were capable of controlling their spending, decision-makers soon turned their attention to control principles developed in the private sector and, consequently, initiated a number of experiments to test the applicability of these instruments in the public domain. The rest is history: while often criticised, NPM has developed into, and today remains, the western hemisphere's dominant public sector management ideology (Arnaboldi, Lapsley, & Steccolini, 2015; Binderkrantz & Christensen, 2009b; Binderkrantz, Holm, & Korsager, 2011; Hyndman & Lapsley, 2016).

The Danish case is no outlier (Greve, 2006; Klaudi Klausen, 2010). Over the recent three decades, the Danish Ministry of Financial Affairs has introduced a broad set of NPM principles and procedures, involving a high degree of centralisation and formal authority to top management of agencies and ministries (Binderkrantz & Christensen, 2009b; Binderkrantz et al., 2011). However, in recent years, NPM has come under fire and has been criticised for shaping a public sector that overemphasizes control and measurement at the expense of service quality and efficiency (Hood & Dixon, 2015a, 2015b). Again, the Danish case confirms the general picture in that a plethora of voices agrees that Danish NPM practices are ridden by flaws and failures (Deloitte, 2011; Devoteam/NextPuzzles, 2011; Kaspersen & Nørgaard, 2015; Kvalitetsudvalget, 2015; Møller, Iversen, & Andersen, 2016).

These concerns, however, have not given rise to redirecting current NPM practices. In fact, recent years have seen the Agency for Modernisation (a Ministerial public agency in charge of the practical implementation of NPM in the Danish public sector) accelerates the deployment of NPM by refining its standards and expanding its range of operation (Agency for Modernisation, 2017). A key element of this acceleration is the use of management accounting schemes that aim to visualise unused resources and redirect them to more cost-efficient initiatives.

In this paper, we analyse a particular NPM tool, namely 'objective and result-based management': a performance measurement system based on a contractual relationship between parties, typically a ministry and a governmental agency, such as a hospital or a university (Agency for Modernisation, 2014b).

Originally the use of performance contracts was initiated as an experiment in the early 1990s; since then it has grown into an almost universal feature of central government (Binderkrantz & Christensen, 2009a, 2009b; Binderkrantz et al., 2011). As pointed out by Ridley and Simon as early as 1938, the effectiveness of such performance contracts is a direct function of their validity (Ridley & Simon, 1938, 1943), which is to say that they must display measures that are controllable and unequivocal and inform agents about the desirable results (Merchant, 1985). Without these features, the likely results are incentive distortion and randomness in

resource allocation and performance evaluation, rendering performance contracts unsuited for strategic planning and management control (Anthony, 1965).

However, given the complex role and organisation of the public sector (Fryer, Antony, & Ogden, 2009; Van de Walle, 2008), there appears to be a whole range of public work areas where it seems impossible to develop controllable and unequivocal measures or even define measurable strategic objectives. In fact, the very definition of good public service is quite often debatable and perspective dependent begging the question of how performance management models may be used in a context where the very concept of public service is open to interpretation.

Following this line of thinking, the current paper sets out to evaluate the validity of the conceptual framework of 'objective and result-based management' used in the Danish public sector. As such, the study is in contrast to traditional management accounting research as it examines how an NPM performance model in itself may contain conceptual deficiencies that are predisposed for problems of validity. To the best of our knowledge, analyses of the internal quality of accounting methods and performance contracts are few and far between and so, this paper seeks to shed light on a perspective that seems underexposed (Baldvinsdottir, Mitchell, & Nørreklit, 2010).

Drawing on pragmatic constructivism, the paper contributes by outlining a theoretical framework stating the requirements for a performance measurement system to display validity in practice; and a method for the analysis of the validity of such a system in a practical context. It subscribes to the notion that performance management systems are human constructs, but not all constructs are equally good at creating functioning activities.

The paper is structured as follows. Section two develops a theoretical framework and method for the analysis and presents the case and contextual background for 'corporate management' in the Danish public sector. In section three, we start off by presenting the contextual analysis of the Agency for Modernisation and their role in promoting 'corporate management', which is followed by the analysis of the performance contracts between the Ministry of Higher Education and Science and the universities in Denmark. In section four, we discuss our findings from the contextual and the operational analyses. Section five presents our conclusion and a discussion on future directions for performance management in the public sector.

2. Methodology

Analysis of the validity of a management tool requires a definition of the criteria to which a management accounting tool must adhere if we want it to be considered valid. In the next sections, we discuss these criteria leading to our theoretically profound research questions. Also,

we outline the method used for undertaking a theoretically rooted analysis of our case. However, we set out briefly describing the context of 'objective and result-based management'.

2.1 Corporate management in the Danish public sector

NPM is a broad concept that consists of a series of elements (Hood, 1991, 1995). However, one of the most important strands of NPM was the stress on private sector management styles (Hyndman & Lapsley, 2016), which in the Danish case was implemented based on a corporate management philosophy with an embedded control principle called 'ministerial governance'. A key tenet is the notion of the corporation: a collective consisting of a ministry, a government agency and an institution with the ministry as the corporate headquarters, and with the ministry having the overall responsibility for controlling and managing the related agencies and institutions in an effective and efficient manner.

The Ministry of Finance is responsible for the dissemination of corporate management across the entire Danish public sector. To this end, the Agency for Modernisation was established in 2011; its transverse objective was to develop an effective, efficient and flexible public sector with healthy and well-functioning workplaces capable of delivering top-quality welfare to the citizens (Agency for Modernisation, 2014a). One of the key tasks was to promote and facilitate corporate management through guiding and inspirational material, such as theoretical work and case examples. In doing so, the agency developed a performance management model (called the 'objective and result-based management'[15]) that was to be broadly implemented in the Danish public sector. The Agency for Modernisation describes the model like this:

> *Objective and result-based management constitutes the strategic foundation for the control of public institutions; it thus helps ensure that effective institutions deliver results and outcome in accordance with certain political goals... the attention of the public sector must be on pinpointing which results to pursue and how to achieve them... Quite simply, objective and result-based management is about decomposing the strategy into measures and about orchestrating a desirable evaluation concerning the realization of measures ... it is a management task to formulate goals and monitor whether goals are met*
>
> (Agency for Modernisation, 2016)

With the 'objective and result-based management' model, the Agency for Modernisation aimed at creating one of the most effective public sector administrations in Europe (Agency for Modernisation, 2014a). In doing so, strategic planning and management control involved not only the planning and control of costs and revenues but also the identification of slack and the optimisation of scarce resources in the realisation of strategic objectives (Agency for Modernisation, 2014c).

[15] It should be noted that Danish institutions are regulated by legislation and resource allocation, 'object and result-based management' is therefore a phenomenon created to support strategic planning and management control within the institutions and should not be confused with regulation resulting from legislation or resource allocation.

The 'objective and result-based management' model is associated with result control, i.e. a management model which through the introduction of result goals sets up indirect control of employees' activities (Merchant, 1985). As such, result control is well established as a performance management model for the governance of decentralized corporations where principals transfer their decision-making authority to agents. At the core of such principal-agent relationships lies the assumption of an opportunistic agent with more information about alternative actions and their consequences than the principal (the so-called problem of asymmetric information); this constitutes the agency problem of moral hazard, i.e. the question of how the agent is motivated to act in the interest of the principal. As a result, the principal [e.g. a Ministry] placed at a higher organisational level requires accounting information about their agents' performance [e.g. a public institution of the ministry] at the lower organisational levels to ensure that agents abide by their principals' interests. The practice of result control has been developed to ensure that agents are both motivated and held accountable for acting in the interest of the principal. To hold agents accountable for the results of their actions is the 'stewardship function' of accounting, which involves the "monitoring and reporting on the custodianship of resources" (AAA, 1966).

In the governance of private corporations, result control is established through, for instance, ROI or EVA as historically based accounting measures for performance evaluation. Although incomplete, the language of financial accounting is a general language that allows for financial evaluations and comparisons of activities across time and space. However, the public sector has no generic currency that can measure welfare across services (Anthony & Young, 1999). Performance measurement in the public sector is therefore challenged by the inherent difficulty of defining clear-cut objectives of services, which is why objectives are often elusive and ill-defined (Kaplan, 2001; Ridley & Simon, 1938). Defining quantifiable objectives for public organisations constitutes one of the most difficult tasks in the whole field of measurement, yet it is a prerequisite to performance measurement (Ridley & Simon, 1938). Nonetheless, it is one reason why private sector management models are very difficult to transfer successfully to the domain of public organisations (Forbes, 1998; Kaplan, 2001). Instead, the conceptual framework of these models must be reflected and developed in the context of the public sector, requiring that the role of the organisation and the concept of 'good performance' be clearly identified (Van de Walle, 2008).

The focal point of this paper is to study to what degree the Agency for Modernisation has been successful in developing the 'objective and result-based management' model so that it meets basic validity criteria.

2.2 Theoretical framework

Today's performance management literature is dominated by two meta-theoretical strands: realism and social-constructivism (Chua, 1986; Ryan, Scapens, & Theobald, 2002, p. 37). Realism assumes that an independent world exists 'out there' – one that can be observed objectively (Frege, 1879; Whitehead & Russel, 1910-1913). It relates validity to the correspondence theory of truth, which considers truth as a match between statement and observations of the world. In accordance with this view, an objective representational truth of accounting information should be established and used to evaluate performance. However, this view has been criticized by a camp of social constructivists who argue that reality is in fact a construction, reified by human thought, discourse, agreements and interaction (Meyer & Rowan, 1977; Miller & O'Leary, 1987). From this point of view, no objective economic reality exists 'out there', and thus language does not reflect any uncorrupted, neutral facts of Economic Reality (T. Tinker, 1991, p. 298) but rather constitutes them.

While the social constructivist argument has some analytical merit, it also poses the risk of generating a social world in which 'anything goes' (Feyerabend, 1970/2010). If no objective reality exists and if language does not have any representational function, the unfortunate implication is that anything may be posited as the truth about the world. However, as Umberto Eco puts it: reality has 'lines of resistance'. In the real world, anything does not go. For real actors who aspire to develop a well-functioning social world, all actions are not equally feasible and not all descriptions are equally helpful. As we believe social constructivists have not properly accounted for this issue, we argue that there is a need to establish an alternative ontology that paves the way for a meaningful formulation of criteria for valid management accounting models. This ontology should acknowledge, on the one hand, that human actors construe organisational reality but also, on the other hand, take into account that not all human constructions are equally well-functioning. The paradigm of pragmatic constructivism has been developed in an effort to do just this.

Pragmatic constructivism

Pragmatic constructivism is built on the core assumptions that human beings are creative and reflective actors who use language to construct activities and coordinate actions in an effort to build functioning practices (Jakobsen, Johansson, & Nørreklit, 2011; H. Nørreklit, Nørreklit, & Mitchell, 2016). A key tenet is that human practices are organised around the use of language games where thoughts, actions and language are interwoven into a totality (Wittgenstein, 1953, §7). A classic example of a language game is 'builder's language' where a builder needs his assistant to bring him various building materials. For this purpose, they develop a language that consists of the words 'block', 'pillar', 'slab', 'beam': the builder calls them out, and the assistant brings the stones he has learnt to bring at such-and-such a call (Wittgenstein, 1953, §2). In other

words, when exclaiming 'block!' the builder does not simply point to the physical block; he actually asks his assistant to bring him a block.

The same applies for the world of accounting. For instance, when planning a restaurant menu, the managers involved need to develop a particular language (e.g. consisting of concepts such as contribution margin, price, costs, targets, estimates, number of customers, product mix, main dish, starters, desserts, etc. (Ahrens & Chapman, 2007, p. 13); subsequently they may interact to collect, calculate and evaluate information about customer behaviour and restaurant activities and hence compile the restaurant menu (i.e. life form). In this sense, language is not external to the world; on the contrary, it is used as a toolbox by actors to construe and develop particular ways of life.

However, in contrast to a radical social constructivist position, pragmatic constructivism emphasises that what actors say does not linearly translate into a well-functioning reality. The success of actors in building a set of effective functioning actions requires the integration of four dimensions of reality: facts, possibilities, values and communication. First, a factual basis for a possible thinkable idea of a life form must be in place. To extend the restaurant example from above, a set of facts is required to make recognised new possibilities of a restaurant menu actually up and running. For instance, the restaurant managers need access to kitchen facilities, ingredients, employees, financial resources, etc., all of which constitute the factual basis for the actual construction of the outlined idea of a restaurant menu. If none of these facts are in place, the possibility of a restaurant menu (or any other social reality) as an end result is an illusion. Furthermore, from the point of view of pragmatic constructivism, it is crucial that the array of factual possibilities somehow reflects the actors' values. If an action cannot be interpreted as meaningful or valuable, actors are disinclined to carry out the action in question. Finally, if actors are to build a well-functioning reality, they must communicate in order to construct and coordinate functioning actions.

As already mentioned, pragmatic constructivism is inspired by the Wittgensteinian idea that the meaning of language is learned and developed in local practices. Hence the meaning of any concept is to be explained by the role it plays in the actors' construction of a functioning practice. Nevertheless, it emphasises that effective communication requires a linguistic structure where each concept has a fairly well-defined and shared meaning. In this paper, we suggest that a concept must the following four criteria of meaning:

First, a concept must be given an abstract meaning. Drawing on their linguistic toolboxes, actors must outline the abstract idea of a concept by defining its cognitive content. For instance, a block is a brick composed of clay or cement that can be used for building houses; 'variable costs' are costs that change in total in proportion to changes in manufacturing volume over short periods of time, etc. This enables actors to point to exemplary references of the concept, which

is useful when it comes to differentiating between that which is and that which is not characterised by the concept.

However, to establish a shared horizon of understanding of what the abstract idea implies in its practical use, actors need to agree on a specific set of *exemplary references*. For instance, in a specific practice of accounting, actors might agree that variable costs signify payments to individual employees for piecework and raw material costs. Mismatches between the abstract ideas and the exemplars may lead to states of confusion and illusion (H. Nørreklit, Nørreklit, Mitchell, & Bjørnenak, 2012), for instance, if you outline an abstract idea of variable manufacturing costs and point to the salary of the director of manufacturing.

Nevertheless, relating a concept to an abstract idea might result in too broad definitions which will be inadequate for planning and control purposes. Therefore, a concept must also be given a supplementary criterion-based meaning. Criteria have the ability to overcome subjectivity issues by transforming the qualitative basis of the conceptual content into numbers. For instance, by having building blocks categorised by size, a construction manager is able to make more detailed planning and control of a building practice; or a segregation of cash flows with a detailed time dimension, might enable managers to engage in more effective financial planning. Importantly, the criteria should be chosen in accordance with the pragmatic use of the concept in question. Thus, a concept with loose criteria may be useful in contexts where individuals' actions are based on intuitive judgment as, for instance, in the arts, while a tighter use of criteria is required when actions are to be externally restricted as, for instance, in science. However, the linking of the conceptual content with conceptual criteria of quantification can be a hub for producing illusions. For instance, in a social context driven by strong forces for quantification, one might put numbers on phenomena that cannot be counted.

Finally, concepts should not only point to things from their appearance but also the aforementioned four dimensions of reality construction should be reflected as layers in the conceptual content. Similar to the need for the integration of the dimensions for a reality construct to function, these layers must account for the concept so as not to produce illusions and confusion. For instance, it is not sufficient to point to numbers of available blocks, we also have to know what the possibilities of their features are for the construction work, and are these features of value for the building projects. Similarly, an account in our bookkeeping may look like an account receivable, but whether it is actually an account receivable depends on whether or not an actual debtor has received some products or services and that they can and will actually make the payment.

The establishment of these qualities in the structure of the concepts is crucial for the validity of the actors' language use. If the concepts are in place, the model in question may assist the actor to facilitate the building of functioning practices. The complexity of the layers and the changeability of the reality construct itself charge the concepts with requests for change and

development; this creates the risk of creating and applying fuzzy, inconsistent illusion-creating concepts, which may in turn be a cause of problems in the organisation. But for concept to be considered useful in the management of a specific practice, it must display a fairly stable conceptual structure. If not, it becomes a 'floating signifier' (Levi-Strauss, 1950/1987); a sign that has the potential to signify all kinds of meanings and therefore is useless in actors' actual construction of their realities.

Nevertheless, the central issue is whether or not the conceptual model facilitates actors' creation of a functioning practice, or, in pragmatic constructivist lingo, whether or not the model allows actors to build 'construct causality' (H. Nørreklit et al., 2012; L. Nørreklit, 2017). Thus, the final test of effectively constructed concepts is related to whether or not the expectations created by the conceptual model are fulfilled. Accordingly, it is only through its application in actual practice that a conceptual model may be tested. Importantly, the degree of truth and reliability of the concepts applied is not an a priori issue but a purely pragmatic one. If the concepts work, i.e. the expectations created by the communication are fulfilled, the statements are true. The pragmatic test can be made by continuously comparing the proactive true claims of the conceptual framework regarding expected outcome of action, and the pragmatic truth as the actual result of action (H. Nørreklit, Nørreklit, & Mitchell, 2007).

As mentioned in the introduction, we know that making the new public management models work as promised in the local practices is problematic, or, in other words, they do not meet the test of being pragmatic true (Kaspersen & Nørgaard, 2015; Møller et al., 2016). In view of that, the question that will be discussed in the remainder of this paper is whether or not the 'objective and result-based management' scheme lives up to the requirements to the structure of the concepts. If this is not the case, the objective and result-based management' model is liable to produce illusions rather than functioning organisational practices.

2.3 Research questions and method

On the basis of the above description of the empirical context and theoretical reflections, we may now raise our research questions:

RQ1: What characterises the conceptual qualities embedded in the model of objective and result-based management?

RQ2: Given the pragmatic constructivist definition of validity as described above, do these conceptual qualities signify a valid model that may facilitate construct causality?

To answer these questions, we analyse the concepts used in the presentation of the performance contract. The methodology used is conceptual enquiry (H. Nørreklit, 2017; Wilson, 1969; Wittgenstein, 1953). A conceptual enquiry is aimed at investigating and evaluating the quality level in the structuring of the concepts used in the models. Drawing on the characteristics of a functioning language game outlined above and by analysing the dimensions of abstract

meaning, criterion-based meaning and exemplar, we look at how the concepts acquire meaning. Also, we analyse the extent to which the concepts integrate the four dimensions of reality and reflect on the conceptual framework's ability to contribute to the actors' construction of a functioning practice.

The analysis reveals that the conceptual structure of the model is ripe with poorly outlined concepts and mismatches allowing a language game of illusions. This finding leads to our third research question:

RQ3: How may we explain that a language game of illusion exists in the realms of performance management of the Danish public sector?

In order to answer this research question, we reflect on the language features from the point of view of a stream of literature rooted in critical sociological theories. This literature reveals that the implementation of performance management models might be explained from some ideological views of installing a social order rather than the more functionalist view of creating functioning practices.

2.4 Selection of case and case material

Analytically, we set out by examining the 'objective and result-based management' at a contextual level by both describing the main terms in the guidelines and inspirational material developed by the Agency for Modernisation and by analysing the conceptual qualities. Furthermore, by looking at the contractual relationship between the Ministry of Higher Education and Science and the seven Danish universities for the years 2015-2017, we analyse how the 'objective and result-based management' model unfolds. Again, we both describe the main terms in use and analyse their conceptual qualities.

These analyses, in conjunction with the theoretical framework of pragmatic constructivism, are used to study the validity of the contractual relationship between the Ministry of Higher Education and Science and the universities. This approach should allow us to shed light on the continuous existence of validity problems related the ministerial governance of the Danish public sector, and hence whether the performance measurement approach is liable to produce illusions rather than functioning organisational practices.

We chose the case of Higher Education and Science because it provides an actual implementation of corporate management and thus may open the window to a rich understanding of a 'real world situation' (Otley & Berry, 1994). In addition, this case is relevant as a study object as it reflects an area where the implementation of performance measurement is known to be a highly difficult and complex task (Arnaboldi et al., 2015). Therefore, the case may reflect a situation of intensity where the reasoning of corporate management is highly prevalent.

Our archive of data consists of qualitative material. The primary data is the material produced by the Agency for Modernisation on corporate management in Denmark along with

the actual contracts between the Ministry of Higher Education and Science and the Danish universities for the period 2015-2017. Our secondary data consists of external reports on both the application of corporate management and the contractual relationship between the ministry and universities as well as newspaper articles on the NPM situation in Denmark.

3. Analysis of 'objective and result-based management'

The current section begins by analysing the conceptual foundation of the performance management model 'objective and result-based management' developed and diffused by the Agency for Modernisation. The result gives rise to two concerns pertaining to the validity of the model. Subsequently, we examine to which degree these concerns are addressed in the actual implementation of the model, namely in the contractual relationship between the Ministry of Higher Education and the Danish universities (i.e. corporation of Higher Education and their subsidiaries).

3.2 The management control approach of the Agency for Modernisation

According to the Agency for Modernisation, 'objective and result-based management' is formulated with the purpose of constituting the strategic foundation for the control of public institutions thus contributing to ensure effective institutions which deliver results and effects in accordance with the political goals.

The Agency's approach to performance measurement and management is encapsulated in Figure 1. The model consists of three interlinked phases, namely 1) the development of strategic objectives, 2) the formulation of measures and targets, and 3) an evaluation phase. In the model, we see the contour of a planning and feedback loop known from conventional wisdom of management control systems (e.g. the *management system* (Kaplan & Norton, 2008, p. 8).

In the following, we will analyse the validity of the model's conceptual content, mainly by analysing the vast amount of guiding material developed by the Agency for Modernisation clarifying the model and also exemplifying ideal ways of actualising the model. We will only focus on the first two phases of the model, as the third phase does not offer insight for the scope of our investigation.

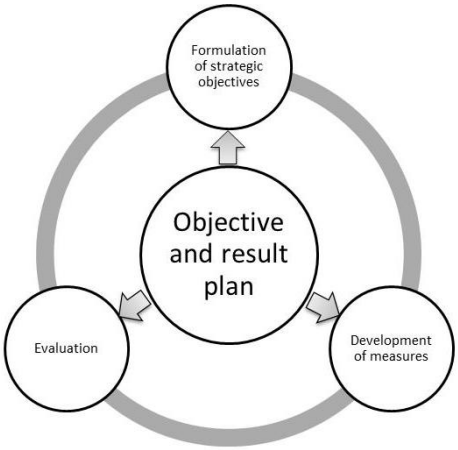


Fig. 1. 'Objective and result-based management' model

Formulation of strategic objectives

The first step of the model is the formulation of strategic objectives. In the 'objective and result-based management' model, the formulation of strategic objectives is defined as a brief description of the strategic targets that are expected to saturate the agency's operations in a multi-year period. Thus:

> *The strategic objectives summarize the agency's strategic prioritizations for a multi-year period...the strategic objectives should be the coherent story of what the agency wants to change and how the agency adds value to its surroundings. In brief, it should be seen as the agency's aim for effectiveness*
>
> (Agency for Modernisation, 2014a, p.15).

Formulating a *'coherent story of what the agency wants to change and how the agency adds value to its surroundings express'* is a fairly vague outline of the concept of strategic objectives and the activities it involves. One may argue that in the quote, the concept of strategic objective is limited by stating that it should be seen as *the agency's aim for effectiveness*. However, the material does not outline the conceptual content, criteria or exemplar of effectiveness in the public sector. Accordingly, there are many possible ways of formulating strategic objectives representing an aim for effectiveness.

In order to anchor the objectives at the agency top management level, the objectives are supposed to be co-authored in a dialogue between the top management of the ministry and that of the agency:

> *The design of the strategic objectives is to be done on a solid foundation, as a cooperation between ministry and agency and is to be anchored at the top management level so that direction, legitimacy and sufficient ownership is ensured... while takes an offset in the political goals for the agency*
>
> (Agency for Modernisation, 2014a, pp. 15-16)

However, as the strategic objectives should *mirror the political priorities* (Agency for Modernisation, 2014b, p.7), the model indicates that it is the ministry that has the final say as to

which strategic objectives are de facto selected. So, also the hierarchical process of formulating strategic objectives is open for interpretation.

Overall, we find the text describing the process of formulating strategic objectives ambiguous; there is not much conceptualisation of how to formulate these objectives or what constitutes a 'proper' strategic objective. This is problematic, considering that the task of defining measurable objectives is one of the most difficult tasks in the entire field of public sector performance measurement (Ridley & Simon, 1938, 1943) – because what is the role of the public sector? And what is good performance? (Fryer et al., 2009; Van de Walle, 2008)

Development of measures

The second step of 'objective and result-based management' is the development and formulation of concrete measures. Thus, the strategic objectives cannot stand alone; they must be complemented by measurement to make the agency attain the desired strategic objective.

> *The focal point of 'objective and result-based management is concrete measures. The formulation of measures is a pivotal precondition for accomplishing the agency's strategic objectives, as they comprise the constituent parts in the ongoing work to realise the results sought*
>
> (Agency for Modernisation, 2014b, p.8).

Four types of measures within two overall areas are given precedence by the Agency for Modernisation. These four types are *quality, activity, output, and effect* measures (Agency for Modernisation, 2014a) related to either operational tasks such as *production, revenue, task handling time, user satisfaction et cetera* (Agency for Modernisation, 2014b, p8) or policy core tasks such as *preparation of political initiatives, negotiations, analytical work, new legislation et cetera* (Agency for Modernisation, 2014b, p. 8). These are exemplars of measures that are intended to assess whether or not the agency actually delivers the desired results as put forward in the strategic objectives. Strangely, however, the Agency for Modernisation does not seem to be concerned with the conceptual content of the strategic objective being measured and the relationship between the measures. Thus, the extent to which the measures measure what they are intended to measure is neglected.

Instead, when the Agency for Modernisation try to explain how to formulate measures, they focus on the criteria of the measures themselves, arguing that they should be shaped in accordance with the so-called SMART framework (Agency for Modernisation, 2014a, p25) as developed by Doran (1981). The framework is driven by the idea that the world can be measured objectively and that measures thus should be (1) specific, (2) measurable, (3) achievable, (4) realistic (but ambitious), and (5) timely. Accordingly, we witness the outline of the scientific thinking of the natural sciences assuming that there is an objective world out there that can be measured quantitatively.

It is not only surprising but also disturbing that there are no content or criteria for any measurement concepts. For instance, there are no considerations as regards outlining the conceptual content of effectiveness or how it can be measured, and efficiency is not even mentioned, which is unexpected as efficiency and effectiveness are central concepts of public sector performance measurement theory[16] (Anthony, 1965; Anthony & Young, 1999; Barnard, 1938; Ridley & Simon, 1938, 1943).

Furthermore, when top management selects initiatives, presumably these make up antecedents of events that push towards an intended effect (Agency for Modernisation, 2010a pp. 9-24; 2010b, p.3). For example, the Agency for Modernisation defines the quality of a measure as its *"characterization of being unequivocally connected with the effect that is being measured and that the measure is to be clearly influenced by the initiatives an agency sets in motion"* (Agency for Modernisation, 2010b, p. 10). In this light, the key challenge is to select the initiatives that have the strongest and clearest influence on the outcome in the direction desired. Accordingly, these measures are set up to realise the desired ends thus entailing an assumption of a push causal relationship between the proposed initiatives and the intended results[17].

The Agency for Modernisation suggests that the ideal way of achieving this in practice is to develop a so-called 'change theory' (Agency for Modernisation, 2010a). In short, a change theory is an empirically proven chain of causal effects that supports the implementation of a specific initiative. Thus, a change theory is not accomplished simply by describing a logical or coincident relationship between two or more variables (in the material described as a "logical model"); the relationship must be verified causally. Figure 2 illustrates a change theory suggesting that an increased teacher's salary is causally linked to improved pupil performance.

[16] In this relation, we define efficiency as the "*optimum relationship between input and output"* (Anthony, 1965, p. 28) or "*the efficiency of administration is measured by the ratio of the effects actually obtained with the available resource to the maximum effects possible with the available resources"* (Ridley & Simon, 1938, p. 23). While, effectiveness is defined as "*a measurement of the result of an effort or performance indicates the effect of that effort or performance in accomplishing its objective* (Ridley & Simon, 1938, p. 21) or *Effectiveness relates to the accomplishment of the coöperative purpose... When a specific desired end is attained we shall say that the action is 'effective'"* (Barnard, 1938, p. 60).

[17] Push causality is not to be confused with pull causality, where push causality exhibits the features of natural laws while pull causality is a logical inference based on purpose-driven causality. Causal pull is therefore about formulating final causes that may actualize the operating values to create a causal pull driven by purpose and motivation. This would guide managerial effort to formulate motivating values to create a causal pull rather than a deterministic causal push (H. Nørreklit et al., 2012).

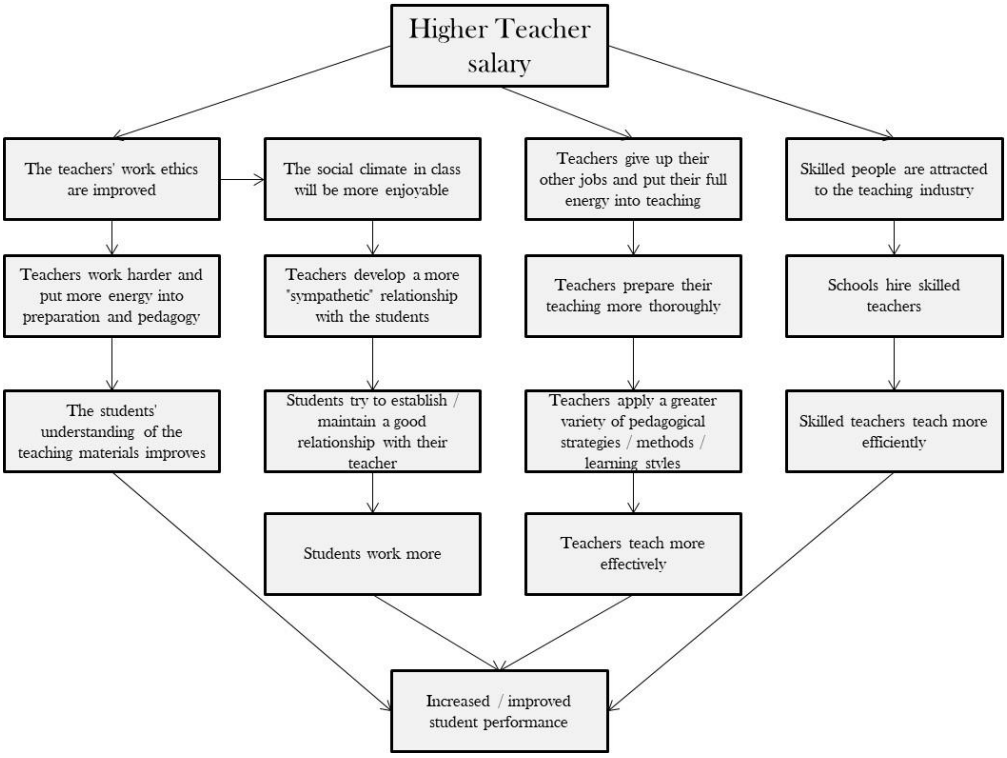


Fig. 2. – A causal schematic on how higher teacher salaries in different ways is causally linked to improved performance of pupils (Agency for Modernisation, 2010a)

In figure 2, we observe a schematic of conditional statements: if event Y, then event X. For instance, if "teachers are payed a higher salary (Y), the objective of improved student performance (X) will be achieved". We also observe that the conditional statements are combined in transitive strings as "higher salary (Y), then teachers' work moral is improved (Z) then teachers will work harder and prepare better (X)" and so on. This type of schematic is built on the assumption of push causality between the different conditional statements and hence something empirically generalizable. This implies that the causal links must be based on sound explanatory theory with empirically verified causalities; otherwise it would be unreliable for the purposes of management control and strategic planning (Anthony, 1965).

Finally, in the material we find that the Agency for Modernisation has defined top management as those in charge of formulating and deciding on measures:

> *Anchoring at top management level is a core factor to the effect and success of the objective and result plan ... Therefore, top management should define and evaluate the content of the objective and result plan ... Otherwise, the strategic management of objectives and results for the most important core tasks will drown in daily operations*
>
> (Agency for Modernisation, 2014a, p. 13).

The quote illustrates hierarchical-top-down-management thinking where knowledge in the local practice is not taken into consideration when formulating measures; also, it illustrates that the Agency for Modernisation regards local involvement as a drawback as it creates the risk

of it drowning in daily operations. On the other hand, it is still the employees *who must deliver the results and meet the targets* (Agency for Modernisation, 2014b, p. 3). Measures must therefore be anchored at the employee level, however, in this context 'anchorage' does not mean to *formulate or decide* on measures (as it was the case with top management); it means that top management should *inform* the employees about the measures.

3.2.1 Conclusion: Two concerns identified

Against this backdrop, we conclude that the aim of 'objective and result-based management' is to ensure effective institutions. At the core of the model is the formulation and deployment of strategic objectives and measures, determined and implemented by top management in a top-down process. Although the formulation of objectives and measures is assumed to take place in a dialogue between a ministry and top management of a governmental agency, the actual interaction process does not indicate any signs of this.

Our analysis shows evidence of the model of 'objective and result-based management' being built on an assumption of a hierarchical-top-down deployment of measures and initiatives that should causally push actions towards the achievement of desired ends. Thus, when management selects a specific initiative for measurement, it is assumed – although with no considerations as to how this takes place at the practical level – to push results through the chain of causality to finally achieve its end. The model seems to follow a mechanistic causal scheme based on the assumption that top managers can push human activity in certain directions thereby accomplishing specific predefined targets. Thus, the model neglects local knowledge about how to establish construct causality and whether it is factually possible to undertake the action that leads to the intended outcome within value range. As such, we believe that it is pertinent to raise two central validity concerns as to the functionality of the model:

First, the measurement terms built into the model seems to be poorly constructed: they are introduced without outlining the cognitive idea and criteria of the concept, and when the guiding material exemplifies the framework, the examples are not linked to the conceptual content. This is illustrated by the vague description of strategic objectives, a description that fails to provide any indications of what is a 'proper' or measurable strategic objective representing a goal for effectiveness. Furthermore, in terms of formulating concrete measures, the focus is put on generic criteria of measures (the SMART framework) rather than on the relationship between measures and the content of the concept being measured. Of course, if this relation is not specified, the actual measures risk being detached from reality and may turn into a system of pure *signifiants* without *signifies,* and hence words without conceptual content. In this case, no guarantee is given that the measures measure what they purport to measure, and thus the model risks stimulating a management practice in which the assumed rudder – the measurements – is unable to induce the aim of establishing effective institutions.

Second, the model's concept of push causality may be problematic when actually put into use. On the face of it, the logic seems impeccable: Top management decides a number of strategic goals and a set of concrete measures that push the organisation towards its desired ends. From the point of view of pragmatic constructivism, it is assumed that there is a deterministic integration of facts and possibility. There are no reflections as to whether construct causality can be established and hence whether the push causality can work. As the model is built on push causality, employees' reflective knowledge about how to establish functioning activities is excluded from the process. In other contexts, employees are considered a rich source of knowledge about *what works*, i.e. what brings the organisation in the desired direction. However, with this type of knowledge out of the equation, the model in question relies unilaterally on external and generic knowledge about possible causal relations between initiatives and outcomes. As a result, if empirical knowledge on cause-and-effect relationships between the transitive strings cannot be produced, the model breaks down and steering becomes rudderless.

In the following, we analyse to which degree these two concerns are addressed in the actual implementation of the model of 'objective and result-based management' in the contractual relations between the Ministry of Higher Education and Science and the Danish universities.

3.3 Analysis of the contract implemented by the Ministry of Higher Education and Science

The Ministry of Higher Education and Science follows the 'objective and result-based management' model outlined by the Agency of Modernisation. The following analysis examines the implementation of the model as regards the two concerns raised before, and we divide the analysis into the same two steps as above. We first look at the strategic objectives formulated by the ministry and then the measures formulated by the universities.

3.3.1 The Ministry of Higher Education and Science

According to the Minister of Higher Education and Science, the performance contracts between the ministry and the universities are formulated with the following purpose:

> *A clear and open dialogue between the ministry and the educational institution about the prioritization of goals, strategies and evaluation while documenting and rendering visible the performance and results of the educational institutions to the outside world*
>
> (Minister, 2015).

The purpose of the performance contracts apparently follows the logic formulated by the Agency of Modernisation for the 'objective and result-based management' model. It is about the achievement of strategic objectives. Specifically, the ministry formulates five strategic areas to be measured: (1) higher quality in education, (2) higher relevance and higher transparency, (3) better coherence and collaboration, (4) strengthened internationalisation, and (5) higher social mobility

– more talents in play. In terms of unit of analysis, we will focus on the strategic objective of 'higher quality in education' and how it is measured.

The strategic objective of 'higher quality in education'

The Ministry of Higher Education and Science wants the universities to strive for higher quality in their educational programs. They base this reasoning on the notion that high quality will permit students to obtain the best possibilities for achieving relevant jobs and will contribute to the creation of new jobs. It is clear that the ministry has a strong focus on the functional aspect of higher education; however, they also argue that education should contribute to the cultivation of students and their active participation in the Danish democracy (Minister, 2015, p. 2). Nevertheless, this latter part has not been conceptualised or specified as to what this means in practice or how it should be obtained or quantified through measures.

The ministry does outline an understanding of what activities they expect would lead to the realization of the functional aspect of 'higher quality in education': *the possibility for students to be full-time students, the development of courses through digitalisation and strong management and strategy on the implementation of digitalisation* (Minister, 2015, p. 2). The minister's briefing describes that these activities are dominated by a causal schematic formulated as a set of conditional statements: *If the universities do (Y) then the objective (X) is achieved.* Sometimes written in transitive strings. Accordingly, it is a reasoning that conforms to the ideas developed by the Agency of Modernisation in the guiding material related to 'objective and result-based management' (Agency for Modernisation, 2010a, 2010b).

A number of examples can be given (for practical purposes the extracts are placed in exhibit 1). The first quote outlines a transitive string of conditional statements: *if* student intensity *then* quality in education, *if* real possibilities of being a full-time student *then* student intensity, and *if* sufficient teaching and guidance hours *then* real possibilities for being a full-time student. The second quote assumes that *if* new digital teaching methods *then* increased study intensity and *if* increased study intensity *then* enhanced quality in teaching and education. The final statement assumes *if* strong management *then* teachers will systematically work with digitalization, create coherent solutions and build up necessary capabilities.

1. "Study intensity and the students are a significant factor in creating quality in education. Educations must be organized so that there are real possibilities for being a full-time student... the individual educational institution must ... support this. Teaching must be prioritized and it is necessary to ensure that the extent of teaching and guidance hours are sufficient"
2) Digitalization in teaching can help enhance quality in teaching and education – New teaching methods can support increased study intensity and internationalization.
3) Strong management and strategic anchoring is crucial for spreading and integrating digitalization... systematic work among teachers and management... and to create coherent solutions and build up necessary capabilities

Exhibit 1. Examples of a causal schematic constructed as transitive strings

The statements in exhibit 1 reflect the measurement logic promoted by the Agency of Modernisation (Agency for Modernisation, 2010b, p. 10) in the sense that they assume the existence of a push form of causality between various factors and quality in education. What is observed in these cause-and-effect strings is the notion of a management prescription of *right* actions leading to the desired results, e.g. 'higher quality in education'. This can be interpreted as providing hypothetical imperatives: '*if a manager wants the goal, then he simply has to implement the first antecedent. If he does so, then the goal will necessarily be achieved without further concern for local practice'*. In this case, if full-time accessibility to study programs, digitalization, and strong manager and strategy implementation, then it is *causally certain* that high quality in education will result.

Furthermore, we see that these strings' modalities and adjectives heavily stress action *'significant, must be prioritized, is necessary, et cetera'*. Structure, digitalization and management are what trigger achievement of results, while employees and students appear to be passive. There is no focus on the doing or the how of improving quality in educations; it is assumed as something that occurs when certain initiatives have been taken. Overall, the concepts are broad and ambiguously defined. For example, the quotes are not specific about the type of digitalization and its implications for the specific dimensions of quality that it may influence nor do they provide evidence that these strings are empirically proven causalities.

In addition, instead of providing concrete information on how the strategic objective of 'higher quality in education' can be quantified, the Ministry of Higher Education and Science repeat the Agency of Modernisation by arguing that measures should translate the strategic goals into specific targets for each year of the contract period and that the SMART framework should be followed to align measures with the strategic goal, ambitions, etc. It appears that the measurability of the strategic objective is just assumed. From the briefing on the performance contracts, we find that measurement is not linked to criteria on conceptual reasoning; there is no discussion of what are the aspects that define the measured concept. No reflection on content or criteria of any measurement concepts.

We see that the two concerns we raised for the conceptual foundation of 'objective and result-based management' has so far not been addressed. In the next section, we look at how this influences the construction of performance measures.

3.3.2 Analysis of the university performance contracts

In this section, we analyse the measures formulated in the seven Danish universities'[18] performance contracts for the period 2015-2017 as regards the strategic objective of 'higher quality in education'. Exhibit 2 illustrates all formulated measures. We have coded the measures

[18] Copenhagen University (KU), Aarhus University (AU), Southern University of Denmark (SDU), Aalborg University (AAU), Copenhagen Business School (CBS) and Technical University of Denmark (DTU)

into the following categorizations of measurements: 'A' *Student efficiency*, 'B' *Program content, C 'Teaching capacity', 'D' Student dedication,* 'E' *Student satisfaction*, 'G' *Digitalization*, and 'F' *Pedagogical skills*.

In the following, we only present our analysis for Copenhagen University, Copenhagen Business School and the Technical University of Denmark; this is done in order to reduce the amount of redundancy in analysis. After a walkthrough of each of the three performance contracts, we present a discussion and conclusion on their conceptual qualities and relationship with the strategic objective of 'high quality in education'.

Danish University Performance Contracts, 2015-2017

Objective:	*Higher quality in education*					
Aarhus University (AU)	**Copenhagen University (KU)**	**University of Southern Denmark (SDU)**	**Aalborg University (AAU)**	**Roskilde University (RUC)**	**Copenhagen Business School (CBS)**	**Technical University of Denmark (DTU)**
Key Performance Indicators (KPIs)						
1. Student satisfaction with their educational program (from 88 to 89% over three years) 2. All bachelor programs should as a minimum offer 12 hours educational activity per week 3. Introducing a new shared e-learning platform (Blackboard) • 2015: Blackboard must be available to 95% of all Bachelor's and Master's degree programmes. • 2016: Min 85% of all permanent staff have received an offer to develop their skills in using Blackboard for teaching activities • 2017: Action plan which states which course elements can be rethought and redesigned with Blackboard	1. Increased study commitment. Students are expected to increase their average production of ECTS credits by 5% from 2015-2017 2. Increase in number of short, practice-based courses or development projects providing teaching skills development for lecturers. • 2015: Four courses or projects with an educational consultant are offered • 2016: Six courses or projects with an educational consultant are offered • 2017: Eight courses or projects with an educational consultant are offered	1. A higher ratio of first priority applicants (from 48% to 51% over three years) 2. A lower drop-out rate among students on the bachelor programs (from 15% to 13,5% over three years) 3. Student satisfaction regarding the commencement of the study (from 81% to 84% over three years)	1. Ensure research-based training by reducing the proportion of part-time employed academic staff to a maximum of 7.9% in all three years. 2. To ensure that all AAU study programs are of a high academic standard, AAU will review its program portfolio during the course of the contract period to ensure sustainability and quality in the programs offered 3. Maintain its maximum extension period of the prescribed study completion time of 1.2 months (within a margin of three months)	1. Consolidating the offer of combination educations (So that 60% of all students enroll in one of these programs in 2015, and in 2017 the target is 80%). 2. Strengthening the RUC learning tools: (1) Create an academic study environment that motivates the students to participate in learning activities. (2) Project competencies and management training. (3) Training in academic reading, writing and critical reflection 3. A lower drop-out rate among bachelor students (from 19% to 16% over three years)	1. Increase student satisfaction with full-time programs. • Measured as an average and weighted scale of overall satisfaction and satisfaction with academic acheivement, teaching, administration and campus environment. 2. Number of courses that are developed online or as blended learning • With at least a 25% increase annually in number of courses. A course is blended when it has an online component of at least 25%.	1. A minimum of 99% of all teaching hours are completed. 2. Digitalization of education through Coursera, which is a platform for online materials that supports active learning processes and functioning as a supplement to lecturing. Measured by counting the number of unique logins from 6800 to 9500 in three years. 3. Student assessment of learning outcome of lecturing is kept at a high level during the contract period (≥3.8)

Exhibit 2. Overview of the 'higher quality in education' performance contract

Copenhagen University (KU)

KU aims to attain 'high quality in education' through two measures formulated within the themes of pedagogical skills (G) and student efficiency (A). KU has developed a measure that looks at the number of short, practice-based courses or projects developed. This activity is argued to increase lecturers' teaching skills; a pilot project has demonstrated an untapped potential for developing lecturers' teaching skills by focusing efforts on their needs and wishes. The performance contracts do no define the content of the courses or projects. KU also measures the percentage increase in students' average production of ECTS credits. This indicates that KU perceives this as a measure that leads to 'high quality in education'.

Copenhagen Business School (CBS)

CBS, on the other hand, formulated measures within the themes of student satisfaction (E) and digitalization (F). CBS measures full-time student satisfaction as a collective student assessment of the scholarly contribution, teaching, administration and student environment. CBS also measures digitalization. This is measured as the introduction of blended learning in courses or as courses developed with online features. The target is an annual increase of 25% in the number of courses in each of the contract years. According to the performance contract, CBS expects this to boost the quality of the study programs and consequently the students' grades.

Technical University of Denmark (DTU)

DTU has formulated three measures within the themes of student satisfaction (E), digitalization (F) and teaching capacity (C). DTU measures student satisfaction as student assessment of learning outcomes, which they want to keep at or above 3.8 in the contract period. In terms of digitalization, DTU measures this as the number of unique logins to their online teaching platform, 'Coursera'. It is a platform created to support active learning processes and hence stimulate educational quality. Lastly, DTU focuses on teaching capacity, through measurement of the scheduled teaching completion rate, i.e. the degree to which teaching schedules were actually followed.

A discussion of measures in the performance contracts

Overall, we find that none of the universities outline the content of educational quality. However, as Exhibit 2 above indicates, we witness that the ministry and the universities have in fact aimed at operationalising the 'change model' as advocated by the Agency of Modernisation seeing that the measures display logic of push causality. For instance, blended learning or logins to an online learning platform push towards higher quality in education. Furthermore, we observe a significant variety and diversity in the seven universities' approaches to the measurement of activities pushing quality in education. Clearly the measures of activities pushing educational quality are not exemplified in the same way across the seven universities, which indicate some

conceptual confusion as to how quality in educational programs should be understood and quantified. Also, the concrete measures in the contracts are not outlined by content or criteria.

Specifically, when looking at the measures explained in the contracts of KU, CBS and DTU, we find that the content of the push activities is broad, open for interpretation, and hence the effect of the activities is uncertain. For instance, when KU takes the initiative to develop teachers' pedagogical skills through practice-based courses, it might improve educational quality, but it will depend on the content of these courses and project. Just because something is framed as pedagogical improvements, achieving this does not necessarily follow. Similarly, when KU argues that average production of ECTS points is a driver of educational quality, it might suggest something different. For instance, when raising program quality, it may in turn influence program difficulty, which may increase dropout rates or failure rates, implying that higher quality could result in a lower average production of ECTS points.

Likewise, we see the schematic of push causality in regard to measures of digitalization in the DTU and CBS performance contracts. But when DTU measures logins to a digital teaching platform as 'Coursera', no information on the activities of its use or about the learning outcome of using the platform is conveyed. Similarly, CBS argues that by developing courses as blended learning or with online features, they can activate students through quizzes, discussion forums, etc. Again, these activities do not necessarily lead to quality in higher education. In the name of digitalization, it appears that universities might take its value for given. Digitalization is assumed to be something management can implement and then teachers can automatically improve their teaching so that student performance is increased. We acknowledge that digitalization contains many possibilities for enhancing lecturing. However, it requires that digitalization is meaningfully linked to the creation of educational results. Also, different course types require different forms of digitalization; this implies that the local knowledge of the particular practice should be taken into account.

Moreover, CBS and DTU measure student satisfaction, but the rules and procedures for measuring student satisfaction are poorly defined. Thus, measure of satisfaction is subjective and hence may represent diverging views such as a positive take on academically challenging work or getting an easy education. Furthermore, it is questionable whether students can gauge the quality of their education it not being a well-defined phenomenon. Finally, dissatisfaction or satisfaction might be grounded outside the study program frame and therefore has no relation to its quality.

Lastly, CBS focuses on teaching capacity measured as the completion ratio of pre-scheduled lectures. Of course, it is a prerequisite to quality that teaching is actually conducted; however, it should not be considered as a driver of higher educational quality. This could only be considered in the rare case when lectures tend to be cancelled and even then, it would be an unambitious measure as one would expect that planned teaching is done.

A conclusion on the measures in the performance contracts

Consequently, we can conclude that there is a large variety in the formulated measures of which accomplishment should lead to higher quality in educational programs. The linguistic structure of the terms used for the measures are poor. Rarely content or criteria of the measures are outlined. As the content and criteria of quality in education are elusive and unclear, we do not know what the concept signifies. Does it mean a high level of student satisfaction? More teaching hours? More IT systems? A mixture of all of this?

With a clearer understanding of what is meant by quality, we might have found more similarities across the formulation of measures. However, in the current performance contracts, many thinkable measures could have been included, rendering the concept of quality almost meaningless. The formulation of measures appears to be based more on intuition than sound reasoning.

Also, the underlying causal schemata as to why these measures will lead to higher quality are not explained and in most cases, appear questionable. Assuming push causality, a suggested initiative should be linked to empirical verification, but such evidence is not presented. Instead we are left to envisage why and how the fulfilment of these measures should or could lead to improvement in educations.

From a pragmatic constructivist point of view, the assumption of push causality is problematic as it fails to integrate the four dimensions of reality. Similarly, in the natural laws of science, push causality involves a deterministic integration of facts, possibility and values. However, as our analysis of the university case reveals, the subjects of measurement are very broad, and hence an extensive range of factual possibilities is embedded in the measure and only some of them might, in certain educational practices, integrate with the Ministry of Education's intentional values of quality in higher education. We acknowledge that some specific initiatives within the area of measures might contain possibilities for creating improvement in the quality of higher education, but this requires the initiatives to be meaningfully linked to the local practice creating the educational results. But as the initiatives according to the Agency of Modernisation are to be implemented hierarchically top down, local knowledge of the particular practice about how to create functioning activities (construct causality) is not taken into account.

As an end result, the performance contracts become meaningless if viewed from the purpose of result control or management control. Or put differently: If we evaluate quality in educations on the basis of the contract measures, we cannot meaningfully ascertain whether or not more quality is obtained. Nevertheless, the Minister of Higher Education and Science maintains that these measures are a way of documenting the performance and results of Danish universities to the outside world. If the targets of the measures are fulfilled, the ministry would assume that students are to excel within their field, but the measures analysed above are almost without any reference to quality in educations, and it would be an illusion to suggest otherwise.

4. Discussion of findings

4.1 A language game of illusory realism

According to the Agency of Modernisation, the purpose of performance management in the public sector is to increase efficiency and effectiveness in public activities. To this purpose, the agency advocates for the employment of so-called 'objective and result-based management': a management scheme developed to shape effective public institutions to deliver results in accordance with the political objectives. The model follows a mechanistic causal scheme based on the assumption that top managers can push human activities in certain directions thereby accomplishing specific predefined targets. Our analysis of the model gave rise to the following major validity concerns as to its functionality: i) the measurement terms built into the model seem to be poorly constructed; and ii) the model's concept of push causality may be problematic when actually put into use.

Certainly, our analysis displays that when the Ministry of Higher Education and Science initiates 'objective and result-based contracts' with the universities, its contractual material fails to outline the content and criteria of core concepts as, for instance, higher quality in education. Consequently, when the term quality is introduced to university management, it does not inform them about what teaching quality is, it only invokes the intuitive meaning of the concept. In view of that, it might not be surprising that also the analysis of the universities' performance contracts reveals that their underlying concepts are poorly defined. There is indeed an extensive production of measures in the performance contracts that are supposed to lead to higher quality in education, but without outline of conceptual content and criteria, the measures are not valid – rather they found the basis for illusions. In other words, we do not measure what we think we measure.

In addition, our analysis of performance contracts shows that the idea of implementing strings of cause-and-effect relationships, in which some initiatives are expected to push reality in the direction of the desired outcome, are questionable. The construction of intentional results through a schematic of observational activities and causations implies a world-view of mechanical realism. Organisational and human activities are assumed to follow 'natural law-like' causality. However, in the universities' contracts the outlined initiatives leading to educational quality are too vague and uncertain to exist as a law-like cause-and-effect relationship that can make activities push towards desired results. The assumption of push causality does not stand a critical test and thus cannot be considered a means to achieve the results. As a result, the relationships appear to be too subjective which renders the realism illusory.

Also, the contractual assumption of implementing activities of push causality hierarchically top down implies that the contracts are not aligned with knowledge about factual possibilities and values held by employees. Indeed, it is noticeable that managers are perceived to be the only actors capable of developing methods for quality while employees (*in casu* academic scholars) must adapt to existing structures and rules. This appears to be a questionable

approach in a complex and diversified knowledge context such as the university sector. The development of new, more effective methods in practices like these is based on human beings in reflective interaction with each other creating a new type of construct causalities for their practices. This requires knowledge about the specific practice. If the academic scholars' values and knowledge of factual possibilities remain out of sync, it seems likely that employees will lose intrinsic motivation to reflect on new ways of creating construct causality; hence the result will likely turn out to be inadequate in terms of innovation and problem-solving (H. Nørreklit, Nørreklit, & Mitchell, 2010; H. Nørreklit et al., 2012; L. Nørreklit, Nørreklit, & Israelsen, 2006). Accordingly, the idea of push causality implying the standardisation of action into all activities is not effective. One size does not fit all.

The assumption of push causality gives reason to question whether the 'objective and result-based management' approach outlined by the Agency of Modernisation is about result control as the term suggests. Thus, it does not build on the conventional wisdom of result control outlined above, but rather it seems to actualise the principles of action control (Merchant, 1985). Action control ensures that individuals perform the right activities that are supposed to lead to the desired results. The initiatives of push causalities embedded in the contracts are actions defined by the ministry and the university management, which subsequently are to be implemented by management in a hierarchical top down way. The result measures of the framework are therefore not about measuring the performance of quality in higher education, but instead about measuring to which extent the 'right initiative' was implemented. However, the measurement of action is problematic as action control is only effective when sound knowledge about desirable/undesirable actions to achieve results exists and it is combined with the ability to choose the desirable actions.

Overall, the action control embedded in the result contracts of the Danish NPM model differs from the dominating ideas of mainstream decentralized corporate governance of private enterprises. One might argue that we witness a flip of the principal agent model. The assumption seems to be that principals have more information about alternative actions and their consequences than the agent. However, the performance contracts formulated do not illustrate that the management of universities or ministry knows the 'right initiative' leading to the desired result, i.e. higher quality in education. Furthermore, the action control that the ministry opposes on the universities is weak as the prescribed initiatives are vague and open to interpretation. For example, the universities fulfil 82% of all goals and when including goals that are partly fulfilled the ratio increases to 93%.

We therefore find that the stewardship role of accounting is not in place. There is no trustworthy *monitoring and reporting on the custodianship of resources* (AAA, 1966) where principals can hold agents accountable. And hence the material neglects the important management accounting task of solving the agency problem of moral hazard in the context of a

decentralised organisation. In view of the high level of decision authority bestowed on university management to deal with huge amounts of resources, it is problematic that the performance measurement system does not provide information on the stewardship function. In relation to the stewardship function, it is also problematic that the measurement framework does not match the output result of an activity with the resources or effort of accomplishing it (Ridley & Simon, 1938).

Therefore, we conclude that the performance management framework outlined does not live up to basic principles of providing concepts that facilitate the purpose of creating effective public sector institutions; and, when drawing on language features such as causality and measurement, the text ends up creating an illusion of a realist language game. Accordingly, it is questionable whether the implementation of 'objective and result-based management' as outlined by the Agency of Modernisation can lead to efficient and effective public sector organisations. Thus, in their application in actual practice, it is highly unlikely that the conceptual models can meet the pragmatic tests, which was also confirmed by a large study conducted by the Danish Institute for Local and Regional Government Research (Møller et al., 2016).

Given the results of our analysis, it is astonishing that such illusory management concepts have become so widely spread and accepted. We argue that this might be because of the legitimised intention of creating effectiveness in the governance of public sector institutions, might only be present at the surface level. Some groups of people might use the illusory language game more or less intentionally to create a certain social order. In the following, we reflect on the social practice embedded in such illusory language game.

4.2 Why the practice of illusory realism?

NPM has been argued to draw on a management accounting language that purports to shape efficient and cost-effective public organisations (Craig, Amernic, & Tourish, 2014), but that this type of language is delusive as it *is not the language of education or morality or scholarship or learning or community*. The current paper extends this point by showing that the accounting measurement system creates a social space where top managers are not made accountable and where employees are not disposed to develop construct causality. Thus, the paper adds to the growing body of literature that points to some of the flaws and failures of NPM. This begs a crucial question: Given the accumulation of evidence suggesting that some core element of NPM does not work, how may we understand that it continues to dominate the discourse and practice of public management?

While it seems futile to aim at providing any absolute or clear-cut answer to this question, we suggest that it may be fruitful to discuss the question from the point of view of critical management studies (Alvarez, 1998; Alvesson, Bridgman, & Willmott, 2009; Bourguignon, Malleret, & Nørreklit, 2004). From this perspective, management techniques are seen as ideological artefacts that by legitimizing the current social order, serve the interests of the prevalent power structure.

In consequence, management techniques should not be analysed in isolation but must be interpreted as reflections of the set of ideological beliefs, principles and practices that organises and defines organisational reality (Alvarez, 1998; Bourguignon et al., 2004; Mannheim, 1952).

From this point of view, then, the realist language game found in this paper seems to serve as a discursive resource for managers to reinforce and broaden their power. Accordingly, bodies in power need to give a legitimate narrative about the current social structure in order to impose control on their employees (Alvarez, 1998, pp. 28-29; Arnold & Hammond, 1994; Boje, Rosile, Dennehy, & Summers, 1997; A. M. Tinker, Merino, & Neimark, 1982). In this case, the ideology that underpins the performance models seems to provide two narrative elements for this purpose:

First, the cornerstone of the realist language game is the idea that organisations may be controlled and mastered through knowledge about cause-and-effect relationships. If this is believed to be true, it implies that managers are capable of stimulating what Weber (1947) refers to as *rationally purposeful actions*: a feature that imbues managers with a certain type of rational legitimacy. Or in plain English: If employees believe that a manager knows which activities lead to a set of desired ends, the manager in question is endowed with legitimate authority.

Second, result-based performance models communicate that the contribution and achievement of the individual manager are measured through a range of objective and relevant quantitative measurements. Although the validity of the measurements is questionable, the ideological assumption of the result contract is that anyone who works hard will be fairly evaluated and rewarded (Bourguignon et al., 2004). Thus, the result contract communicates that managers are accountable for their actions, which gives rise to what coined *a value-rational legitimacy* (Weber, 1947): a legitimacy based on shared values, in this case the values of accountability and fairness.

Both types of authority very much lend themselves to the "trust in numbers" that dominates western societies (Foucault, 1969; Miller & O'Leary, 1987; Porter, 1996). Numbers are key elements in the scientific genre, which are designed to master and predict the natural world. Accordingly, the use of numbers in management performance models indicates that managers are in control of the uncertainty involved in their jobs, leaving them with enhanced *ethos* and engendering the approval and acceptance of management by their surroundings (Alvarez, 1998, p. 22). In this light, the unclear and flexible concepts found in this paper may be advantageous to managers as the flexibility in meaning makes it easier to hide any lack of knowledge and to pretend that progress has been achieved despite inconclusive evidence.

While the performance management techniques thus further the interests of managers (which is substantiated by the fact that recent years have seen a dramatic increase in managers' salaries relative to that of non-managers (Ammitzbøll, 2011; Klingsey, 2010; Kretzschmer, 2010; Mynster, 2015; Siegumfeldt, 2013), there is also a broader network of actors that seem to benefit from the performance models (Latour, 1987). Most importantly, the preference for the "scientific

approach" built into the models gives primacy to a network of scientific journals, theoretical constructs and individual researchers that favour a quantitative methodology (Merchant, 2010) as opposed to a qualitative one. In consequence of the inherent force in this network, university academics may subsequently adapt to the demands of this network and thus become actors that promote the management models in question. In Staw and Epstein's words, they (or perhaps we should say 'we') contribute to the syndrome of legitimisation (Staw & Epstein, 2000) and thus become active proponents of management by numbers.

This development is problematic in a number of ways, first and foremost because it shapes a university system in which critical thinking and (intellectual) innovation are marginalised (Mejlgaard, Aagaard, & Siune, 2002). While the realist approach gives primacy to university managers' spreadsheets, the autonomy of the individual scholars is challenged as performance models create a direct line of command between politicians, university managers and employees. Essentially, this move away from a classic Humboldian university ideal means that academic scholars become restrained by political and economic interests, restricting the concept of innovation to mean adaptation to the current demands of society and the 'customers'. However, this is an extremely limited concept of innovation failing to acknowledge that true innovation is rarely the result of explicit market demands but rather the spin-off from autonomous researchers' endless experimentation and curiosity. In the current system, critical and curious work that challenges existing concepts, theories and practises is not regarded a virtue; instead researchers are asked to passively adapt to the needs of the market.

5. Conclusion and further directions for research

5.1 Conclusion

Analysing the conceptual qualities in the 'objective and result-based management' approach to the effective governance of Danish institutions, we found the model to be poorly conceptually outlined and with mismatches in the conceptual structure leading to a language game of illusions. Based on the material from the Ministry and the university contracts, it simply was not possible to grasp what perception of quality the universities were striving for or what type of quality the ministry wished them to strive for. In consequence, the measurement appeared to be based on intuition rather than reasoning, which renders the content open for interpretation; in other words, we have no clue as to whether the universities' educational programs are of increasing quality or high quality if we look at the measures in the contracts.

The objective and result-based management model may be sold on the notion of result control, but what we observed was not result control, but action control, which involves the university managers' implementation of right actions in the organisation leading to desired results as delineated by the causal schematics. However, our analysis shows that the causal schematic outlined is obviously too vague, uncertain and general to guide actions that lead to the desired

end. In conclusion, we find that the outlined performance management framework does not live up to the basic principles for providing concepts facilitating the purpose of creating effective public sector institutions.

In view of that, it might not be surprising that in Denmark NPM is seeing continuous problems in creating the intentional effects (Kaspersen & Nørgaard, 2015; Møller et al., 2016). This was confirmed in a large study conducted by the Danish Institute for Local and Regional Government Research concluding that *objective and result-based management works but often not as intended... Many of the factors that appear as inhibitory are about the actual implementation of the objective and result-based management, where it as a theoretical concept cannot 'tolerate' that, for example, the professional judgement is replaced with checklist behaviour or process measures with performance goals.*

However, given the results of our analysis, we found it astonishing that this illusory management concept has become so widely spread and accepted in the Danish public sector. We analysed the language features from the point of view of social theory trying to find a sound explanation of our findings. We argue that that the legitimized intention of creating effectiveness in governance of public sector institutions might only be on the surface level. Some groups of people might use the illusory language game more or less intentionally to create a certain social order. In order to impose a certain control over employees, the bodies in power need to offer a narrative about how to integrate and maintain social order (Alvarez, 1998), and our 'trust in numbers' may render the performance contracts successful and prolific as an instrument for justifying for management authority and control (Foucault, 1969; Miller & O'Leary, 1987; Porter, 1996).

By drawing on elements from the scientific genre, organisations and their managers can show themselves and their surroundings that they are in control of the uncertainty involved in their jobs. Considering this, the unclear and flexible concepts may be an advantage because it makes it easier to hide knowledge gaps and pretend that progress has been achieved for the simple reason that lack of clarity and flexibility make room for whatever explanation at hand. It could therefore be argued that the implementation of NPM and in particular the 'objective and result-based management' framework was bound to fail, as they were never implemented with the purpose of ensuring effectiveness and efficiency in the use of public resources but instead to reinforce the authority of ministries and top management in public institutions and agencies, while showing another narrative to their surroundings.

5.2 Further directions for NPM

Our findings give reason to discuss what can be done to improve performance management of public sector activities? In view of the pragmatic constructivism theoretical framework, there are some basic performance management tasks that NPM has to address in a more conceptually profound way. Immense sums are spent each year by our public institutions and agencies. We

need to develop conceptual tools of performance management to meet the practical need of choosing between alternative courses of action (Ridley & Simon, 1938) and to solve the stewardship task of whether they fulfil the objectives of principals.

However, the paradigm of realism must be abandoned in the control of public sector activities. It is an illusion that the public sector can be controlled through a sort of natural law causality that is founded externally to the local practice. Organisational actors are not part of the natural world and they cannot interpret and control actions mechanically. The functioning of organisational activities in various types of practices, including public sector practices, is based on human beings in local activities organising interwoven strings of construct causalities. The immense complexity of organised constructed strings of causalities is to be managed through leading ideas characterizing the practice. Human beings are creative and reflective constructors of their practices. They should use cognitive abilities to conceptualise and develop an understanding of how to create a functioning practice. Research findings of successful actions may serve as tools of inspiration, but they have to be reflected into the particular practice. Therefore, the role of management is to conduct the process of co-authoring. Mechanical control implies that the person gives up her co-authorship and hands it over to another person thereby giving her authority to determine what to do. If management is reduced to mechanistic control, the likely result is decomposition of sociality (H. Nørreklit, 2017).

Nevertheless, we need to develop concepts to observe the particular purposes of controlling our practices. As institutions of the public sector are not operated for profit, and as the techniques of cost accounting have only limited applicability, we need to devise other criteria for the appraisal of public activities (Ridley & Simon, 1938). If performance measurement is to function as a tool for making intelligent decisions, managers first need to make what public institutions try to accomplish explicitly; and then they must devise methods for measuring the degree of accomplishment. But understanding the concept of quality in higher education depends on whether it is used for the purpose of the Ministry of Higher Education and Science planning and controlling of the university management or of the teachers planning their courses. Also, the quality of the curriculum is not the same as the quality of the teaching method or the quality of the students' study activities. Thus, the meaning of the concept of quality in higher education must be explained by the role it plays in the various organisational actors' creation of construct causalities within their specific activities.

In view of pragmatic constructivism, the intelligent production and use of accounting information are challenging because accounting information is uncertain and partial and hence unlikely to capture the full complexity of organisational reality. Specifically, when talking about a complex phenomenon like quality of higher education, it can be difficult to set up quantitative criteria. Nevertheless, it is essential that there is a conceptual framework of performance management that fulfils basic criteria for conceptual qualities as described above. Thus, the

content, criteria and exemplar of the concepts should be coherently outlined, and the four dimensions of reality construction should be reflected as layers in the conceptual content.

Moreover, we need a *conceptual narrative* for explaining how certain measures should be linked to actions (factual possibilities), objectives (values) and relevant actors (communication). Thus, the measures cannot be linked mechanically to actions but have to be linked to a reflective narrative telling how to create construct causality. Analysing whether all four dimensions of reality are integrated, the validity of a conceptual narrative can be evaluated proactively. However, it is only through pragmatic use that we can detect whether the measurement and the narrative express reality facilitating functioning practice. The practical validity of the measurement narrative is always linked to whether the future action holds true. If the expected results are realised in action, then the statement is pragmatically true. This cannot be observed mechanically but requires that the performance measures must be used in combination with an interactive reflective method investigating whether the performance system generates the pragmatic truth knowledge in its complexity. Also, it implies that the use of measures requires the exercise of professional judgement.

All in all, we need further research on the development of such conceptual frameworks in public sector activities. In particular, it seems fruitful both to investigate the more successful existing practices of performance management in public sector organisations and to do more interventionist research on how to develop more valid conceptual frameworks in particular public sector practices. Finally, we seem to be trapped in the language game of realism while we need new types of language games that might enable us to transgress the realism culture producing illusions.

References

Ahrens, Thomas, & Chapman, Christopher S. (2007). Management accounting as practice. *Accounting, organizations and society, 32*(1), 1-27.

Alvarez, José Luis. (1998). *The diffusion and consumption of business knowledge*. Houndsmills: MacMillan.

Alvesson, Mats, Bridgman, Todd, & Willmott, Hugh. (2009). *The Handbook of Critical Management Studies*. New York: Oxford University Press.

Ammitzbøll, Lisbeth. (2011). Ledere stiger mest i løn. *Magisterbladet*.

Anthony, Robert N. (1965). *Planning and Control Systems. A Framework for Analysis*. Boston: Graduate School of Business Administration, Harvard University.

Anthony, Robert N., & Young, David W. (1999). *Management control in nonprofit organizations* (Vol. 6): Irwin Homewood, IL.

Arnaboldi, Michela, Lapsley, Irvine, & Steccolini, Ileana. (2015). Performance Management in the Public Sector: The Ultimate Challenge. *Financial Accountability & Management, 31*(1), 1-22.

Arnold, Patricia, & Hammond, Theresa. (1994). The role of accounting in ideological conflict: lessons from the South African divestment movement. *Accounting, Organizations and Society, 19*(2), 111-126.

Baldvinsdottir, Gudrun, Mitchell, Falconer, & Nørreklit, Hanne. (2010). Issues in the relationship between theory and practice in management accounting. *Management Accounting Research, 21*(2), 79-82.

Barnard, Chester Irving. (1938). *The functions of the executive* (Vol. 11). Cambridge: Harvard university press.

Binderkrantz, Anne Skorkjær, & Christensen, Jørgen Grønnegaard. (2009a). Delegation without agency loss? The use of performance contracts in Danish central government. *Governance, 22*(2), 263-293.

Binderkrantz, Anne Skorkjær, & Christensen, Jørgen Grønnegaard. (2009b). Governing Danish agencies by contract: from negotiated freedom to the shadow of hierarchy. *Journal of public policy, 29*(01), 55-78.

Binderkrantz, Anne Skorkjær, Holm, Mogens, & Korsager, Kirstine. (2011). Performance contracts and goal attainment in government agencies. *International Public Management Journal, 14*(4), 445-463.

Boje, David M, Rosile, Grace Ann, Dennehy, Robert, & Summers, Debra J. (1997). Restorying reengineering: Some deconstructions and postmodern alternatives. *Communication Research, 24*(6), 631-668.

Bourguignon, Annick, Malleret, Véronique, & Nørreklit, Hanne. (2004). The American balanced scorecard versus the French tableau de bord: the ideological dimension. *Management accounting research, 15*(2), 107-134.

Chua, Wai Fong. (1986). Radical developments in accounting thought. *Accounting review*, 601-632.

Craig, Russell, Amernic, Joel, & Tourish, Dennis. (2014). Perverse audit culture and accountability of the modern public university. *Financial Accountability & Management, 30*(1), 1-24.

Deloitte. (2011). Kortlægning af økonomi- og virksomhedsstyring i udvalgte statslige institutioner København: Finansministeriet.

Devoteam/NextPuzzles. (2011). Analyse af økonomi- og virksomhedsstyring i udvalgte statslige institutioner. København: Finansministeriet.

Doran, G. T. (1981). There's a SMART Way to Write Management's Goals and Objectives *Management Review* (Vol. 70, pp. 35-36.): American Management Association.

Feyerabend, Paul. (1970/2010). *Against Method: Outline of an Anarchist Theory of Knowledge*. London: Verso.

Forbes, Daniel P. (1998). Measuring the unmeasurable: Empirical studies of nonprofit organization effectiveness from 1977 to 1997. *Nonprofit and voluntary sector quarterly, 27*(2), 183-202.

Foucault, Michel. (1969). *The Archaeology of Knowledge* (A. M. S. Smith, Trans. Vol. 2002). London: Routledge.

Frege, G. (1879). *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (H. Nebert, Trans.). Halle: Nebert.

Fryer, Karen, Antony, Jiju, & Ogden, Susan. (2009). Performance management in the public sector. *International Journal of Public Sector Management, 22*(6), 478-498.

Greve, Carsten. (2006). Public management reform in Denmark. *Public management review, 8*(1), 161-169.

Hood, Christopher. (1991). A public management for all seasons? *Public administration, 69*(1), 3-19.

Hood, Christopher. (1995). The "New Public Management" in the 1980s: variations on a theme. *Accounting, organizations and society, 20*(2-3), 93-109.

Hood, Christopher, & Dixon, Ruth. (2015a). *A government that worked better and cost less?: Evaluating three decades of reform and change in UK central Government*: OUP Oxford.

Hood, Christopher, & Dixon, Ruth. (2015b). What we have to show for 30 years of new public management: Higher costs, more complaints. *Governance, 28*(3), 265-267.
Hyndman, Noel, & Lapsley, Irvine. (2016). New Public Management: The Story Continues. *Financial Accountability & Management, 32*(4), 385-408.
Jakobsen, Morten, Johansson, Inga-Lill, & Nørreklit, Hanne (Eds.). (2011). *An actor's approach to management : conceptual framework and company practises* (1. edition ed.). Copenhagen: DJØF.
Kaplan, Robert S. (2001). Strategic performance measurement and management in nonprofit organizations. *Nonprofit management and Leadership, 11*(3), 353-370.
Kaplan, Robert S, & Norton, David P. (2008). *The execution premium: Linking strategy to operations for competitive advantage*: Harvard Business Press.
Kaspersen, Lars Bo, & Nørgaard, Jan. (2015). *Ledelseskrise i konkurrencestaten* (1. udgave ed.). Kbh.: Hans Reitzel.
Klaudi Klausen, Kurt. (2010). Koncernledelse i det offentlige - nu også i kommunerne? *Ledelse & erhvervsøkonomi, Årg. 74, nr. 2 (2010)*, 7-24.
Klingsey, Mette. (2010). Løngabet mellem ledelse og ansatte på KU stiger, *Information*.
Kretzschmer, Liv Alfast. (2010). Fodboldlønninger og tung administration dræner universiteterne, *Magisterbladet*.
Kvalitetsudvalget. (2015). *Nye veje og høje mål : Kvalitetsudvalgets samlede forslag til reform af de videregående uddannelser*. Kbh.: Udvalg for Kvalitet og Relevans i de Videregående Uddannelser.
Latour, Bruno. (1987). *Science in action: How to follow scientists and engineers through society*: Harvard university press.
Levi-Strauss, Claude. (1950/1987). *Introduction to Marcel Mauss*. London: Routledge.
Mannheim, K. (1952). Conservative thought. In P. Kecskemeti (Ed.), *Essays in Sociology and Social Psychology*. London: Routledge & Kegan Paul.
Mejlgaard, Niels, Aagaard, Kaare, & Siune, Karen. (2002). *Politik Og Forskning: Forskningspolitik Mellem Autonomi Og Heteronomi*: Analyseinstitut for Forskning.
Merchant, Kenneth A. (1985). *Control in business organization*: Financial Times/Prentice Hall.
Merchant, Kenneth A. (2010). Paradigms in accounting research: A view from North America. *Management Accounting Research, 21*(2), 116-120.
Meyer, John W, & Rowan, Brian. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American journal of sociology, 83*(2), 340-363.
Miller, Peter, & O'Leary, Ted. (1987). Accounting and the construction of the governable person. *Accounting, Organizations and Society, 12*(3), 235-265.
Minister. (2015). Briefing for contracts. København: Ministry of Higher Education and Science.
Modernisation, Agency for. (2010a). Case-samling – inspiration til effekt: Hvordan vælger du mål, metoder, redskaber og resultatkrav, der passer til jeres styringsbehov? København: Ministry of Finance.
Modernisation, Agency for. (2010b). Ramme for case-samlingen - inspiration til effekt: Hvordan vælger du mål, metoder, redskaber og resultatkrav, der passer til jeres styringsbehov? København: Ministry of Finance.
Modernisation, Agency for. (2014a). About us. Retrieved 19-09, 2019, from http://www.modst.dk/Servicemenu/In-English
Modernisation, Agency for. (2014b). Inspiration til strategisk styring med resultater i fokus. København: Ministry of Finance.
Modernisation, Agency for. (2014c). Strategisk styring med resultater i fokus. København: Ministry of Finance.
Modernisation, Agency for. (2014d). Vejledning: Økonomistyring i staten, del 1 Målbillede. København: Ministry of Finance.
Modernisation, Agency for. (2016). Mål- og resultatstyring. Retrieved 19-09, 2017, from http://www.modst.dk/God-okonomistyring/God-oekonomistyring/Maal-og-resultatstyring
Modernisation, Agency for. (2017). Mål og resultatplan. København: Ministry of Finance.
Mynster, Ivan. (2015). Markante lønstigninger til offentlige topchefer. *Ugebrevet 4A*.

Møller, Marie Østergaard , Iversen, Katrine , & Andersen, Vibeke Normann (2016). Review af resultatbaseret styring. København: KORA.

Nørreklit, Hanne. (2017). *A Philosophy of Management Accounting: A Pragmatic Constructivist Approach*. New York: Routledge.

Nørreklit, Hanne, Nørreklit, Lennart, & Mitchell, Falconer. (2007). Theoretical conditions for validity in accounting performance measurement. *Business performance measurement: unifying theories and integration practice, 2nd edn. Cambridge University Press, Cambridge*, 179-217.

Nørreklit, Hanne, Nørreklit, Lennart, & Mitchell, Falconer. (2010). Towards a paradigmatic foundation for accounting practice. *Accounting, Auditing & Accountability Journal, 23*(6), 733-758.

Nørreklit, Hanne, Nørreklit, Lennart, & Mitchell, Falconer. (2016). Understanding practice generalisation–opening the research/practice gap. *Qualitative Research in Accounting & Management, 13*(3), 278-302.

Nørreklit, Hanne, Nørreklit, Lennart, Mitchell, Falconer, & Bjørnenak, Trond. (2012). The rise of the balanced scorecard! Relevance regained? *Journal of Accounting & Organizational Change, 8*(4), 490-510.

Nørreklit, Lennart. (2017). Paradigm of Pragmatic Constructivism. In H. Nørreklit (Ed.), *A philosophy of management accounting: A pragmatic constructivist approach* (pp. 21-94). London: Routledge.

Nørreklit, Lennart, Nørreklit, Hanne, & Israelsen, Poul. (2006). The validity of management control topoi: towards constructivist pragmatism. *Management Accounting Research, 17*(1), 42-71.

Otley, D. T., & Berry, A. J. (1994). Case study research in management accounting and control. *Management Accounting Research, 5*(1), 45-65.

Porter, Theodore M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*: Princeton University Press.

Ridley, Clarence E, & Simon, Herbert A. (1938). The criterion of efficiency. *The Annals of the American Academy of Political and Social Science, 199*(1), 20-25.

Ridley, Clarence E, & Simon, Herbert A. (1943). *Measuring municipal activities: A survey of suggested criteria and reporting forms for appraising administration*: International city managers' Association.

Ryan, Robert, Scapens, Robert W, & Theobald, Michael. (2002). Research methods and methodology in accounting and finance. *Thomson, London*.

Siegumfeldt, Pernille. (2013). Et lektorproletariat rykker tættere på, *Magisterbladet*.

Staw, Barry M, & Epstein, Lisa D. (2000). What bandwagons bring: Effects of popular management techniques on corporate performance, reputation, and CEO pay. *Administrative Science Quarterly, 45*(3), 523-556.

Tinker, Anthony M, Merino, Barbara D, & Neimark, Marilyn Dale. (1982). The normative origins of positive theories: ideology and accounting thought. *Accounting, Organizations and Society, 7*(2), 167-200.

Tinker, Tony. (1991). The accountant as partisan. *Accounting, Organizations and Society, 16*(3), 297-310.

Van de Walle, Steven. (2008). Comparing the performance of national public sectors: conceptual problems. *International Journal of Productivity and Performance Management, 57*(4), 329-338.

Weber, Max. (1947). *The Theory of Social and Economic Organization*. New York: Oxford University Press.

Whitehead, AR, & Russel, BB. (1910-1913). *Principia Mathematica*.

Wilson, S. (1969). *Thinking with Concepts*. Cambridge UK: Cambridge University Press.

Wittgenstein, L. (1953). *Philosophical Investigations* (G. E. M. Anscombe, Trans.). Oxford: Basil Blackwell.

AAA. (1966). *A statement of basic accounting theory*. Evanston, USA: American Accounting Association.

6. Conclusion: a journey into contemporary performance measurement

Non-financial measures were introduced to performance measurement theory to break with the historical nature of financial measurement (Eccles, 1991; Johnson & Kaplan, 1989; Lueg & Nørreklit, 2013). The concern was that the construction of financial measures meant that they revealed a great deal about past actions but very little about the future. Financial measures do not emphasise any element leading to good or poor *future* financial results, as they are unable to encapsulate uncompleted chains of actions that extent beyond the time of measurement (Nørreklit, 2000). In contrast, non-financial measures were argued to be future-oriented, as they were argued to be leading indicators of future financial performance through their ability to encapsulate elements of future financial performance.

Kaplan and Norton (1996) were the pioneers of this argument, by conceptualising that outcome measures and performance drivers of outcomes were linked together in cause-and-effect. This claim resulted in empirical postulates of generic actions that were to drive successful business performance and they became embedded into contemporary PMSs (Lueg & Nørreklit, 2013). A common example of such a postulate, is the assumption that customer satisfaction is a *certain* driver of future financial performance, and by building PMS on the notion of causation, it is claimed to equip organisations with a tool to *a prior* know what drives the future financial performance (De Haas & Kleingeld, 1999).

However, there exist no conclusive evidence for the existence of these claimed cause-and-effect relations and the *actual* consequences of implementing contemporary PMS in either private or public organisations to some extent remain unknown (Arnaboldi, Lapsley, & Steccolini, 2015; Franco-Santos, Lucianetti, & Bourne, 2012; Hoque, 2014; Micheli & Mari, 2014). As such, the relevance of non-financial measurement as equal to financial measurement stands on a flimsy foundation. What if non-financial measures cannot be found to be a causal driving force of future financial performance? What does this mean for the construction of contemporary PMS and the use of non-financial measures along with financial measures? And, what does this mean to performance measurement in public sector organisations, as financial measurements do not convey any information on the organisational performance of such organisations?

These are important questions that this dissertation has tried to provide some answers to and, next, we will outline the conclusion, contribution and practical implication of each of the four papers in relation to contemporary performance measurement and causality. The last part of the section provides a short summative conclusion and a final reflection of the entire dissertation.

6.1. Conclusion, contribution and practical implications

6.1.1 Chapter 2 (first paper)

This paper aimed at analysing the empirical foundation for the assumption of causality in contemporary performance measurement through providing an answer to the following research question: *How far has PMAR come in providing consistent empirical answers to causal questions(s) of non-financial measures being leading indicators of future financial performance?* To provide an answer to the research question, the paper explored the consistency of evidence in PMAR on whether non-financial measures could be considered leading indicators of future financial performance. The empirical data for the paper was the normal science of PMAR and we therefore employed Kuhn's disciplinary matrix in doing a systematic review of the empirical evidence for causations. In particular, the study focused on the *exemplars*, as they represent the most well-known solutions to the puzzles of normal science (Kuhn, 1970) and they therefore represent the collars of the argument of causality.

In conclusion, we found that the empirical evidence for claiming the existence of causations in contemporary performance measurement was inconclusive, as we found the cause-and-effect relationships to be inconsistent, which contradicts the definition of causality (Hume, 1975). We were therefore unable to provide an empirical backing to the inherent assumption of causality, which means that it remains a *brute* fact (Anscombe, 1958; Fahrbach, 2005). This paper contributes first and foremost to the theoretical discussion of causality in management accounting (Balakrishnan & Penno, 2014; Gassen, 2014; Ittner, 2014; Luft & Shields, 2014; Lukka, 2014; Van der Stede, 2014), by systematically reviewing published literature for empirical evidence on causality. As such, we are unable to back the creation of any meta-laws within performance measurement theory. We therefore advice practice to be sceptical when these contemporary PMSs are advocated on the argument of causality and also in relation to the assumption of a connexion between non-financial measures and future financial performance, despite this being widely presumed in performance measurement package research (Kaplan & Norton, 1996).

6.1.2 Chapter 3 (second paper)

The inspiration for this paper lies within a decade long discussion outside of management accounting research on the irreproducibility of positivistic research and to what extent the phenomenon of Questionable Research Practice (QRPs) could be deemed responsible (Banks et al., 2016; Gelman & Loken, 2014; Gigerenzer & Marewski, 2015; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011).

The discussion of QRPs rests on the argument of the phenomenon distorting the hypothetico-deductive method in favour of a researcher's own hypothesis at the risk of increasing the ratio of false-positives (Ioannidis, 2005). The phenomenon of QRPs is of importance to PMAR, as the *sine qua non* method of this research genre is null-hypothesis testing (NHST) (Chua, 1986; Lachmann, Trapp, & Trapp, 2017; Lindsay, 1994; Merchant, 2010) and due to the

ambition of creating complete and valid maps of causations that are to inform practice on what *'works'* (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2003).

This study takes point of departure in analysing published PMAR from 2010 to 2015 for indications of a publication practice that could allow for QRPs to take root. Our concern is that QRPs are found to be widespread within natural and social sciences and if the publication practices of PMAR are unintentionally allowing for QRPs, we would expect QRPs to be present. As such, it is the ambition of the study to provide an answer to the following research question: *How susceptible are the publication practices of PMAR to the phenomenon of QRPs?* By answering this research question, we shed light on a topic, which so far has not received any attention in PMAR; not even in a recent study on the validity of PMAR published in 2017 (see Lachmann et al., 2017).

In conclusion, we find that the current publication practice of PMAR provides space for QRPs to flourish, and we would therefore expect the ratio of false-positives in PMAR to be well above the assumed ratio of 5%. Therefore, it is likely that PMAR is producing research findings when they should not have been produced.

The purpose and ambition of PMAR are to develop reliable and valid causal explanations of management accounting phenomena, in other words, to draw inferences from a sample of specific observations to the general (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2014). PMAR therefore addresses the following type of questions: *"Do differences in environmental or strategic context lead to differences in management control systems? Do certain activities drive overhead costs? Does the adoption of a balanced scorecard system improve performance?"* (Ittner, 2014, p. 545). The answer to such questions becomes unreliable for the purpose of generalising to practise, when the publication environment producing them appears to be infected by the activities of QRPs, as the risk of generalising a false-positive finding would potentially be too high.

In the end, we argue that the dysfunctionality of the current publication system in relation to QRPs has created the space for a bad equilibrium to unfold and that this equilibrium is sustained due to the 'publish or perish' pressure and the fact that QRPs are a viable solution to insignificant findings. The trigger of the bad equilibrium is arguably the "journals", so for any solution to be a viable solution, this is where the solution needs to be found.

In the paper, we propose three different solutions for the journals to implement, and we also judge how realistic each of these solutions are and if they impact the practice of QRPs enough for it to potentially dissolve. However, if PMAR continues the current trajectory it is likely that production of research findings, if generalised to practice, is going to have unforeseen consequences to the society we as researchers strive to service.

6.1.3 Chapter 4 (third paper)

This paper explored how efficiency and effectiveness criteria relate to the inadequacy of PMS implementations in public sector organisations. The study presents a case study of a PMS implementation in a Danish municipality from the first working document to the latest iteration of the PMS. In the paper, we argue that a measurement of these two criteria must be central for public sector PMS to achieve the strategic objectives with efficient resource consumption under financial constraints. The efficiency and effectiveness criteria render the resource flow from costs, through outputs, to outcome transparent and manageable. In this way, performance measurement is considered as a tool enabling public managers to answer the following questions: (1) *'how adequate and effective is our service performance?',* and (2) *'how efficient are we in providing these services?'* However, empirical results have evidenced that the implementation of PMS in the public sector is rarely a success, as it has not resulted in the expected improvements in performance, accountability, transparency and quality of services.

In particular, the paper addresses the following research question: *To what extent has the management of the municipality met the criteria of efficiency and effectiveness in their performance contracts between top management and organisational units, and what role does the attainment of the criteria of measurement play in ensuring a functioning PMS?*

We found that notwithstanding the endeavours to develop a well-functioning and successful PMS, the analysis evidences that the PMS fails in accomplishing its purpose of directing, actions and activities toward the achievement of strategic objectives. The PMS is therefore unable to create internal transparency, which results in efficiency and effectiveness not being balanced through the activities of management control. As a result, the PMS becomes nothing more than an administrative burden that provides top management with a false sense of security in the optimisation of scarce resources.

The paper contributes with evidencing the importance of efficiency and effectiveness criteria when implementing and designing a PMS that functions according to its theoretical purpose. An importance that is more significant to the public sector as they are more reliant on non-financial measures due to the limited use of financial performance measurement for public organisations.

This study, indirectly illustrates that organisations cannot simply rely on the existence of a cause-and-effect relationship to exist between non-financial measures and future financial performance or another defined outcome objective. It is not universal laws that create successful local practices or render PMSs that create internal transparency, so that the activities of management control can balance an efficient and effective use of the finite resources available to the managers of public organisations.

6.1.4 Chapter 5 (fourth paper)

In this paper, we explored the NPM tool of 'objective and result-based management' developed by the Agency of Modernisation with the purpose of promoting effective and efficient public institutions. It is a tool developed around the idea of ministries and agencies engaging in contractual binding relationships and we sought out to analyse the validity of this framework. It is a study of how a performance model in itself may contain deficiencies that dispose for problems of validity and then how these validity issues unfold in practice. In particular, we addressed the following three research questions. First, *what characterises the conceptual qualities embedded in the model of 'objective and result-based management?* Second, *given the pragmatic constructivist definition of validity as described above, do these conceptual qualities signify a valid model that may stimulate construct causality?* Finally, *how may we explain that a language game of illusion exists in the realms of performance management of the Danish public sector?*

The analysis is divided into two sections. The first is an analysis of the theoretical underpinning of the model, as outlined in the material developed by the Agency of Modernisation. The second section is a study of its implementation between the Ministry of Higher Education and Science and seven Danish universities. We address a need for a better understanding on why the implementation of such NPM tools appear to be continuously failing in the Danish public sector.

Our analysis evidences that 'objective and result-based management' is poorly outlined and with mismatches in its conceptual structure that leads to a language game of illusions. These issues are then translated into the practical application, as the performance contracts between the Ministry of Higher Education and Science and the Danish universities suffers from the same issues. The objective of the contracts was partly to increase educational quality, but it was not possible to grasp what perception of quality the universities were striving towards or towards which type of quality the ministry wished them to strive. This resulted in a highly diverse formulation of measures, that appeared to be based more on impressions than rational reasoning, which rendered the content open for interpretation; in other words, we have no clue whether the university educations are of an increasing quality or high quality if we solely look at the KPIs in the contract. In the end, we found the outlined performance management framework of 'objective and result-based management' to not live up to the basic principles for providing concepts that can facilitate the purpose of creating effective public sector institutions.

This study shows how the practice of founding performance measurement on causal schematics, which involves management prescriptions of right actions leading to desired results failed when implemented in a complex local context such as a university setting. Our analysis showed that the causal schematic outlined was obviously too vague, uncertain and general to guide actions that lead to the desired end.

In terms of the last research question, we speculate if the legitimized intention of creating

effectiveness in governance of the public sector institutions, might only be on the surface level. Some groups of people might use the illusionary language game more or less intentional to create a certain social order. And by drawing on elements from the scientific genre, organizations and their managers can show themselves and their surroundings that they are in control of the uncertainty involved in their jobs. Considering this, the unclear and flexible concepts may be an advantage, because it makes it easier to hide lack of knowledge and pretend that progress has been achieved, for the simple reason that the lack of clarity and the flexibility make room for whatever explanation at hand.

It could therefore be argued that the implementation of NPM and in particular the 'objective and result-based management' framework inevitably would fail, as they were never to be implemented with the purpose of ensuring efficiency and effectiveness in the use of public resources but instead cement the authority of ministries and top management in public institutions and agencies, while showing another narrative to their surroundings.

6.2. Concluding remarks

By questioning the argument of empirical postulates of generic actions that are to drive successful business performance or in another terminology cause-and-effect relations, we have attempted to address the lack of theoretical profoundness concerning the presumption of causality. In doing so, we approached the problematic from two aspects, first, by looking into the empirical grounding of these claims i.e. the identification of 'true' causations, and, second, by analysing contemporary performance measurement implementations in the Danish public sector pioneered by the NPM movement. The dissertation thus not only contributes to the theoretical understanding of causality but also to the actual implementations of PMSs which are constituted by non-financial measures in the anticipation of these measures being drivers of organisational performance. The dissertation evidence that the claim of causality stands on a flimsy foundation as we found no evidence on the existence of true universal causations that could guide and define PMSs implementations in practice. Nothing indicates the existence of eternal and universal relationships between non-financial measures and financial performance.

However, we may be able to establish statistical associations between contingent variables and aspects of practice, however, the methods by which and the reasons why contingencies influence (or do not influence) practice are evidently not well explained (Mitchell, 2017). Pragmatic constructivism can in conjunction with another theory such as performance measurement theory, provide a complementary approach to enable insight to be gained on how and why organisations succeed in creating successful PMSs. As such, we argue that the relationship that exists between non-financial measures and financial performance is something temporarily and unstable that is continuously being constructed by the human actors of the

organisation and we therefore need another framework to study this than the one imported from natural sciences.

References

Anscombe, G. E. M. (1958). On brute facts. *Analysis*, 69-72.

Arnaboldi, M., Lapsley, I., & Steccolini, I. (2015). Performance management in the public sector: The ultimate challenge. *Financial Accountability & Management, 31*(1), 1-22.

Balakrishnan, R., & Penno, M. (2014). Causality in the context of analytical models and numerical experiments. *Accounting, organizations and society, 39*(7), 531-534.

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. (2016). Questions About Questionable Research Practices in the Field of Management A Guest Commentary. *Journal of Management, 42*(1), 5-20.

Chua, W. F. (1986). Radical developments in accounting thought. *Accounting review*, 601-632.

De Haas, M., & Kleingeld, A. (1999). Multilevel design of performance measurement systems: enhancing strategic dialogue throughout the organization. *Management Accounting Research, 10*(3), 233-261.

Eccles, R. (1991). The performance measurennent manifesto. *Harvard business review, 69*(1), 131-137.

Fahrbach, L. (2005). Understanding Brute Facts. *An International Journal for Epistemology, Methodology and Philosophy of Science, 145*(3), 449-466. doi:10.1007/s11229-005-6200-7

Franco-Santos, M., Lucianetti, L., & Bourne, M. (2012). Contemporary performance measurement systems: A review of their consequences and a framework for research. *Management Accounting Research, 23*(2), 79-119.

Gassen, J. (2014). Causal inference in empirical archival financial accounting research. *Accounting, organizations and society, 39*(7), 535-544.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate Science The Idol of a Universal Method for Scientific Inference. *Journal of Management, 41*(2), 421-440.

Hoque, Z. (2014). 20 years of studies on the balanced scorecard: Trends, accomplishments, gaps and opportunities for future research. *The British Accounting Review, 46*(1), 33-59.

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals* (3. ed., repr. / with text rev. and notes by P.H. Nidditch ed.). Oxford: Clarendon.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

Ittner, C. D. (2014). Strengthening causal inferences in positivist field studies. *Accounting, organizations and society, 39*(7), 545-549.

Johnson, H. T., & Kaplan, R. S. (1989). *Relevance lost : the rise and fall of management accounting* (2. print ed.). Boston, Mass.: Harvard Business School Press.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: translating strategy into action*: Harvard Business Press.

Kuhn, T. S. (1970). The structure of scientific revolutions, International Encyclopedia of Unified Science, vol. 2, no. 2: Chicago: The University of Chicago Press.

Lachmann, M., Trapp, I., & Trapp, R. (2017). Diversity and validity in positivist management accounting research—A longitudinal perspective over four decades. *Management Accounting Research*.

Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research, 11*(1), 33.

Lueg, R., & Nørreklit, H. (2013). Performance measurement systems - beyond generic actions. In F. Mitchell, H. Nørreklit, & M. Jakobsen (Eds.), *The Routledge Companion to Cost Management* (pp. 342-359). Abingdon, Oxon: Routledge.

Luft, J., & Shields, M. D. (2003). Mapping management accounting: graphics and guidelines for theory-consistent empirical research. *Accounting, organizations and society, 28*(2), 169-249.

Luft, J., & Shields, M. D. (2014). Subjectivity in developing and validating causal explanations in positivist accounting research. *Accounting, organizations and society, 39*(7), 550-558.

Lukka, K. (2014). Exploring the possibilities for causal explanation in interpretive research. *Accounting, organizations and society*.

Merchant, K. A. (2010). Paradigms in accounting research: A view from North America. *Management Accounting Research, 21*(2), 116-120.

Micheli, P., & Mari, L. (2014). The theory and practice of performance measurement. *Management Accounting Research, 25*(2), 147-156.

Mitchell, F. (2017). A pragmatic constructivist approach to studying difference and change in management accounting practices. In H. Nørreklit (Ed.), *A philosophy of management accounting: A pragmatic constructivist approach*. London: Routledge.

Nørreklit, H. (2000). The balance on the balanced scorecard a critical analysis of some of its assumptions. *Management Accounting Research, 11*(1), 65-88.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 0956797611417632.

Van der Stede, W. A. (2014). A manipulationist view of causality in cross-sectional survey research. *Accounting, organizations and society, 39*(7), 567-574.

7. CO-AUTHOR STATEMENTS

AARHUS BSS SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Declaration of co-authorship*

Full name of the PhD student: Kristian Mohr Røge

This declaration concerns the following article/manuscript:

Title:	A STUDY ON THE CRITERIA OF INTERNAL TRANSPARENCY, EFFICIENCY AND EFFECTIVENESS IN MEASURING LOCAL GOVERNMENT PERFORMANCE
Authors:	Kristian Mohr Røge and Niels Joseph Lennon

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: Financial Accountability & Management

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Major contribution
C. Equal contribution
D. Minor contribution
E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	C
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	B
6. Finalization of the manuscript and submission	C

Signatures of the co-authors

Date	Name	Signature
17-10-2017	Kristian Mohr Røge	
17-10-2017	Niels Joseph Lennon	

In case of further co-authors please attach appendix

Date: 17-10-2017

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.

AARHUS BSS SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Declaration of co-authorship*

Full name of the PhD student: Kristian Mohr Røge

This declaration concerns the following article/manuscript:

Title:	THE ILLUSION OF 'OBJECTIVE AND RESULT-BASED MANAGEMENT': BEYOND AN NPM TOOL IN DENMARK
Authors:	Kristian Mohr Røge, Nikolaj Kure and Hanne Nørreklit

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: British Accounting Review

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Major contribution
C. Equal contribution
D. Minor contribution
E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	B
2. Planning of the experiments/methodology design and development	B
3. Involvement in the experimental work/clinical studies/data collection	B
4. Interpretation of the results	B
5. Writing of the first draft of the manuscript	C
6. Finalization of the manuscript and submission	C

Signatures of the co-authors

Date	Name	Signature
17-10-2017	Kristian Mohr Røge	Kristian Røge
17-10-2017	Nikolaj Kure	NKure
17-10-2017	Hanne Nørreklit	

In case of further co-authors please attach appendix

Date: 17-10-2017

Kristian Røge
Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.